An Introduction to Interdomain Routing and the Border Gateway Protocol (BGP)

Timothy G. Griffin

AT&T Research

griffin@research.att.com

http://www.research.att.com/~griffin

http://www.research.att.com/~griffin/interdomain.html

ICNP

PARIS

November 12, 2002

Outline

1. The Internet is implemented with a diverse set of physical networks
2. Relationships between Autonomous Routing Domains (ARDs)
3. BGP as a means of implementing and maintaining relationships between ARDs
4. BGP as means of implementing local optimizations (“Traffic Engineering”)
5. What Problem is BGP Solving anyway?
6. Current Internet Growth Trends
7. Selected References

In Memory of Abha Ahuja

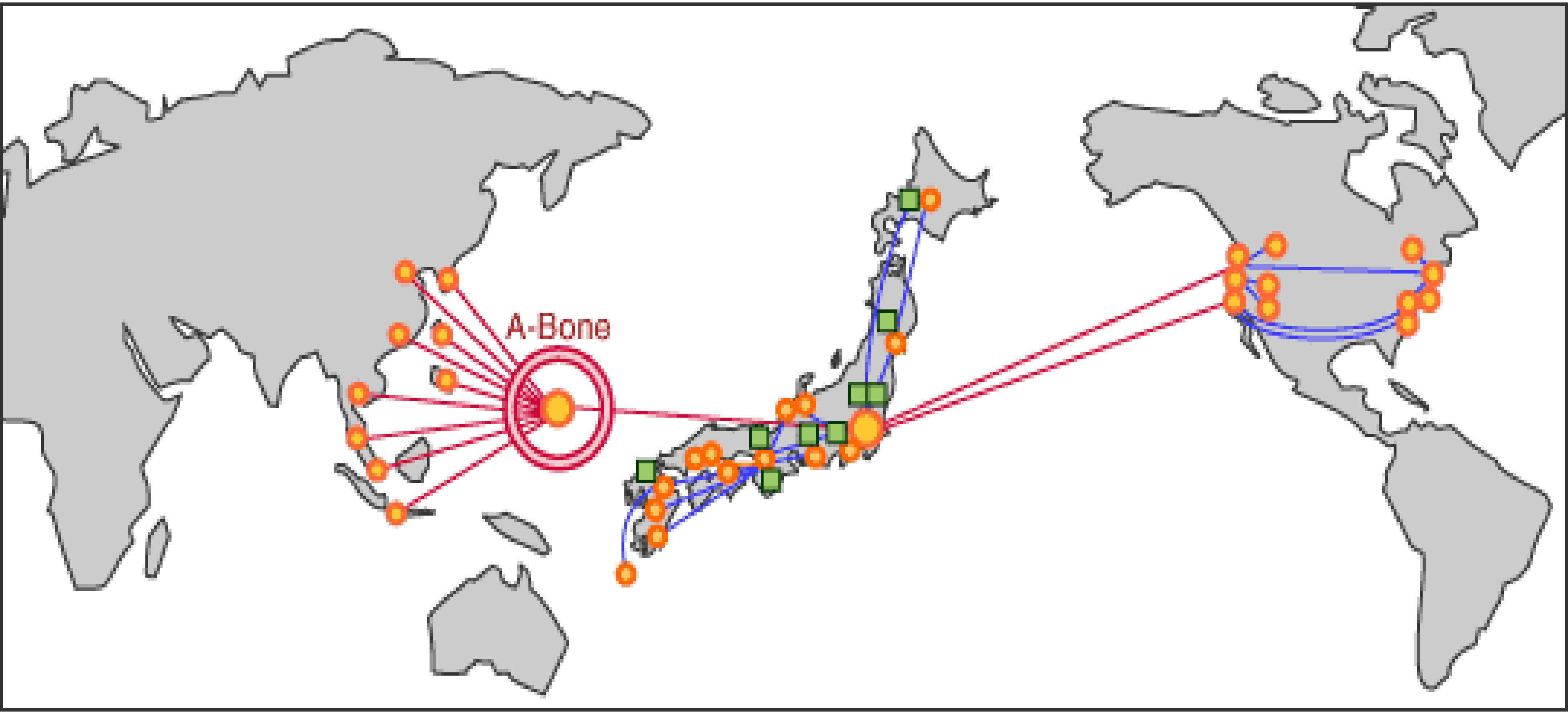
Photo by Peter Lothberg. http://www.caida.org/~kc/abha/gallery.html

NANOG memorial site: http://www.nanog.org/abha.html

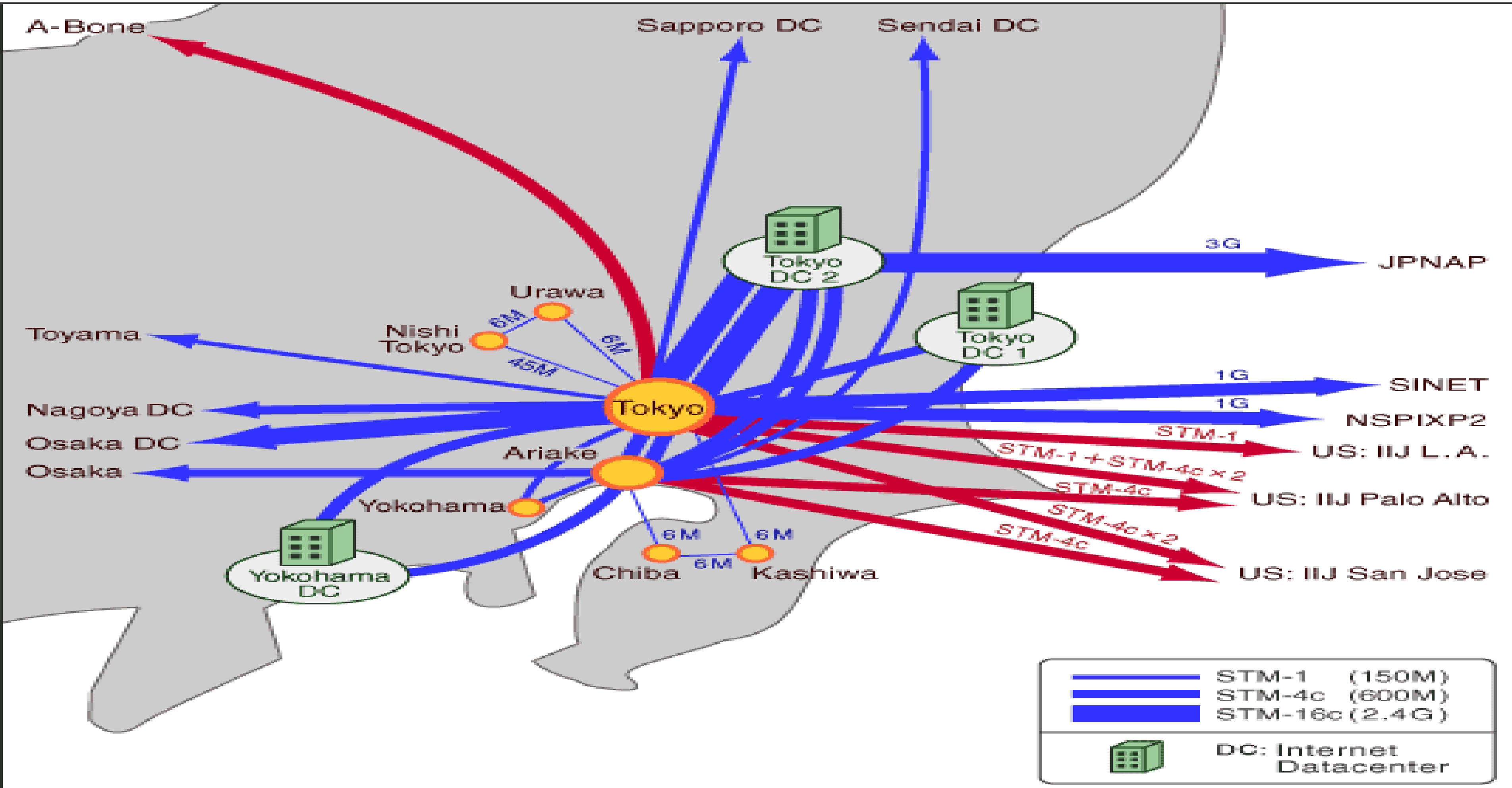
PART I

Physical Connectivity

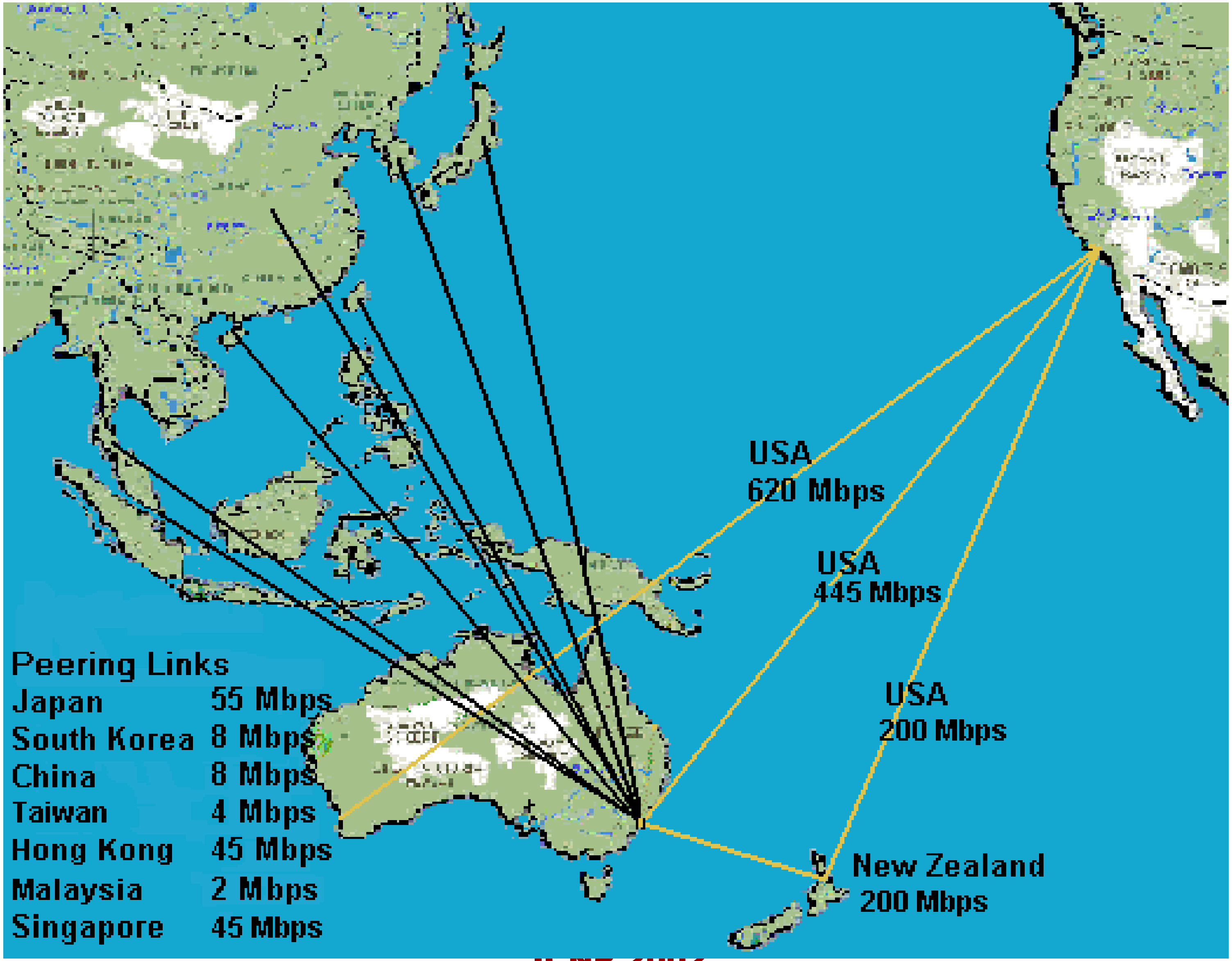
Internet Initiative Japan (IIJ)



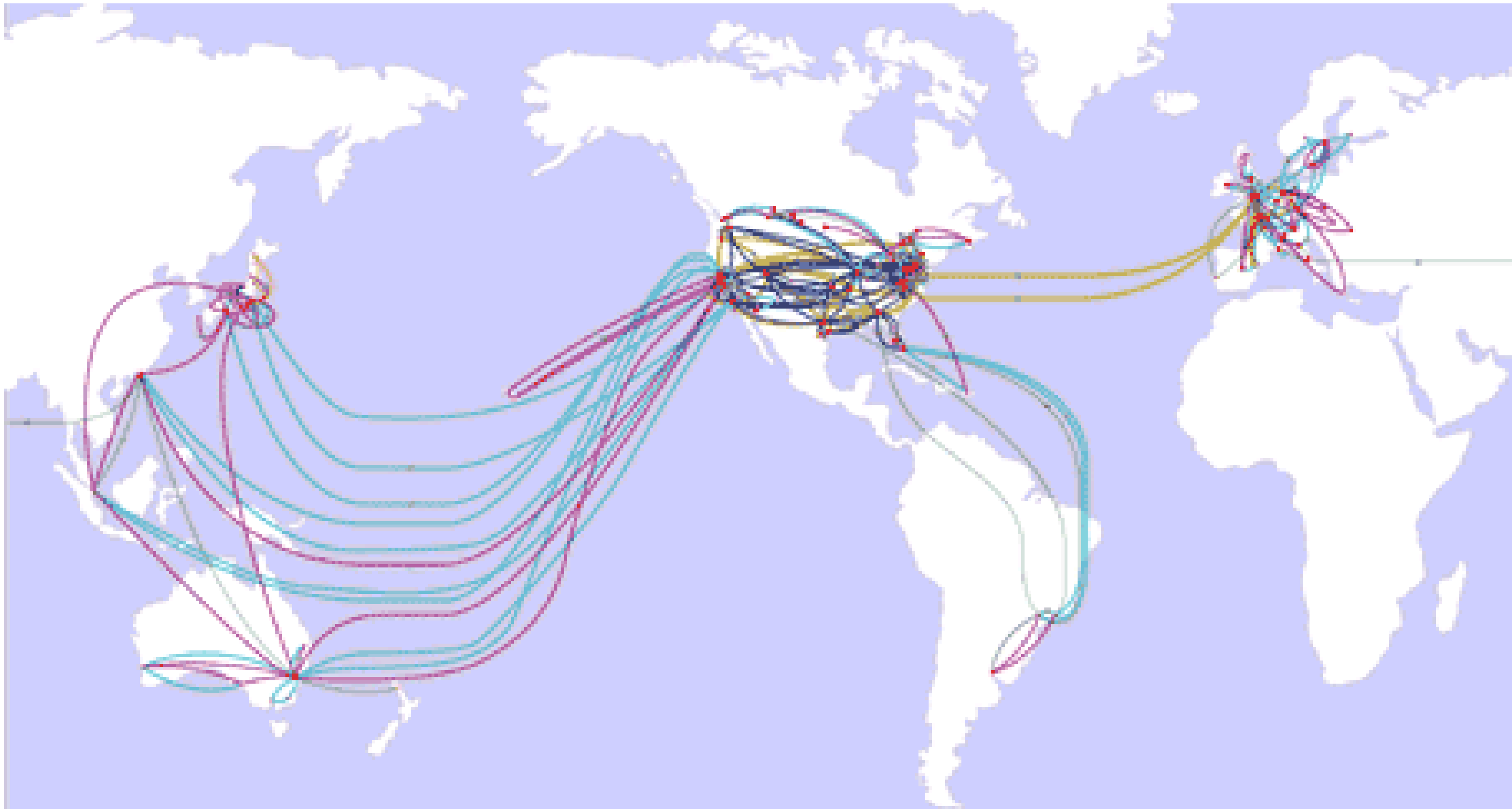
IIJ, Tokyo


A-Bone
Sapporo DC
Sendai DC
Tokyo DC 2
3G
JPNAP
Tokyo DC 1
Urawa
6M
Nishi Tokyo
6M
45M
Toyama
Tokyo
1G
SINET
1G
NSPIXP2
Nagoya DC
Osaka DC
Osaka
Ariake
STM-1
US: IIJ L.A.
STM-1 + STM-4c × 2
STM-4c
US: IIJ Palo Alto
STM-4c × 2
STM-4c
US: IIJ San Jose
Yokohama
6 M
6 M
Chiba
6 M
Kashiwa
Yokohama DC
STM-1 (150M)
STM-4c (600M)
STM-16c (2.4G)
DC: Internet Datacenter

Telstra international

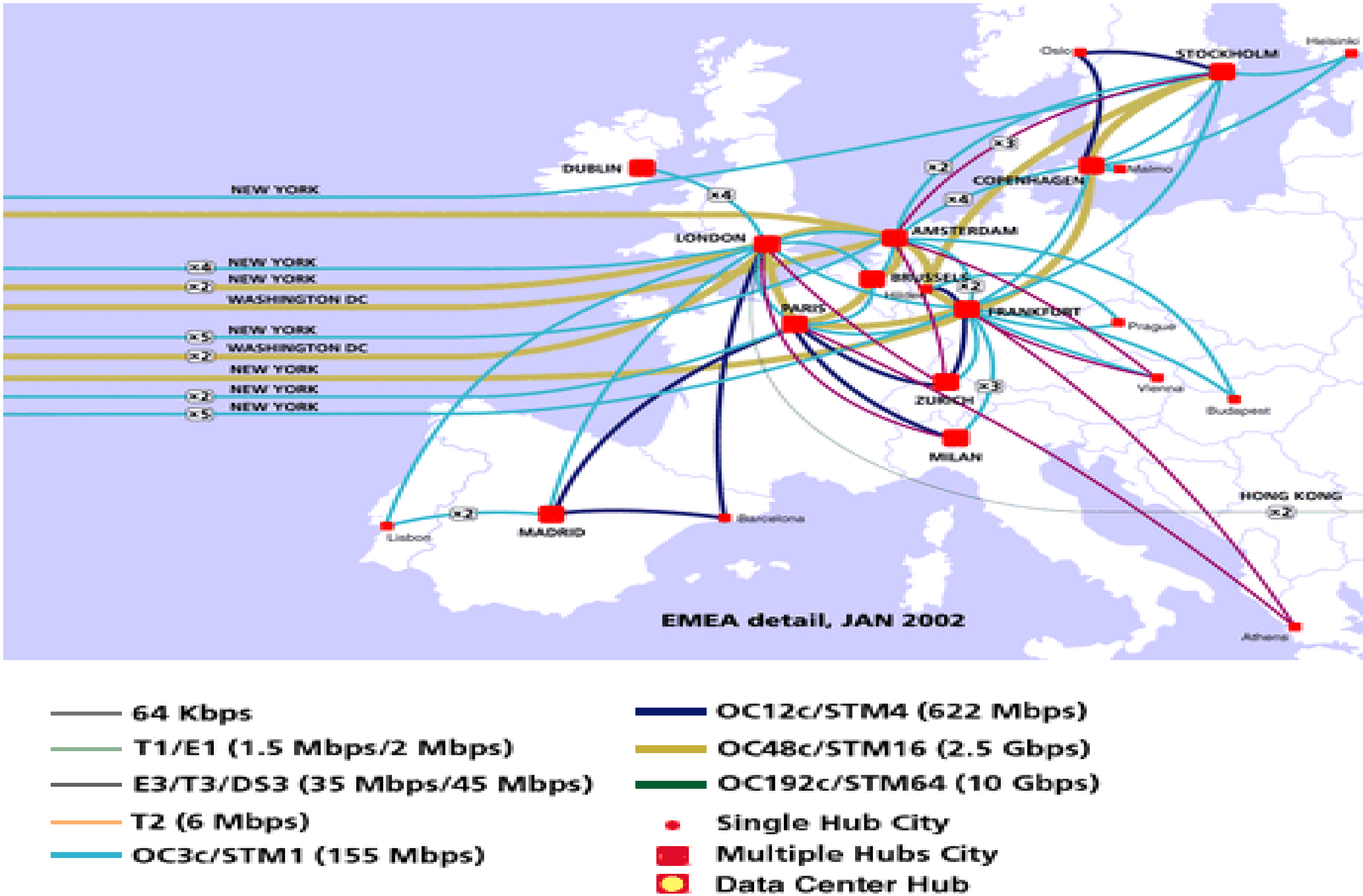


WorldCom (UUNet)

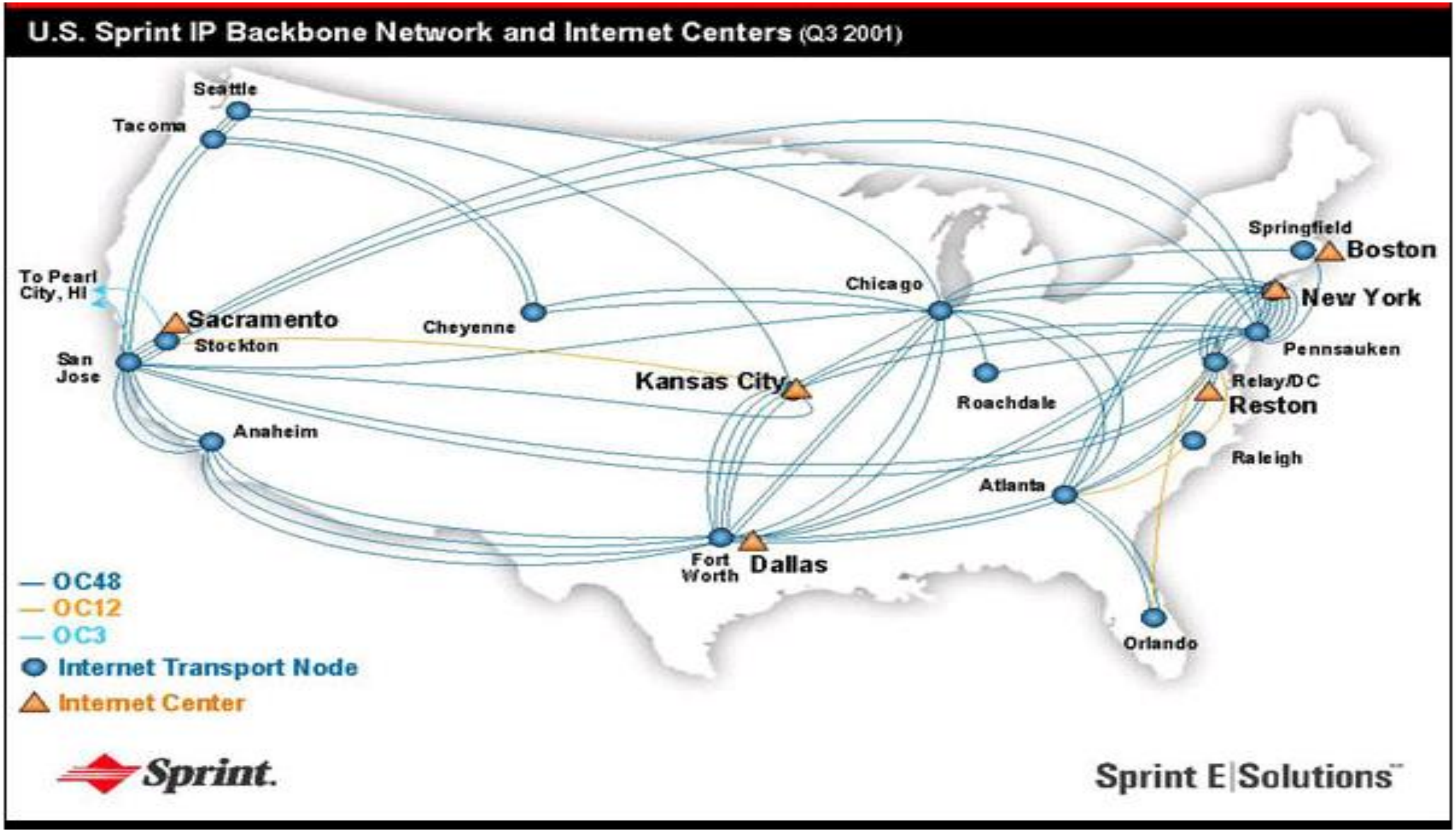


- 64 Kbps
- T1/E1 (1.5 Mbps/2 Mbps)
- E3/T3/DS3 (35 Mbps/45 Mbps)
- T2 (6 Mbps)
- OC3c/STM1 (155 Mbps)
- OC12c/STM4 (622 Mbps)
- OC48c/STM16 (2.5 Gbps)
- OC192c/STM64 (10 Gbps)
- Single Hub City
- Multiple Hubs City
- Data Center Hub

UUNet, Europe



Sprint, USA

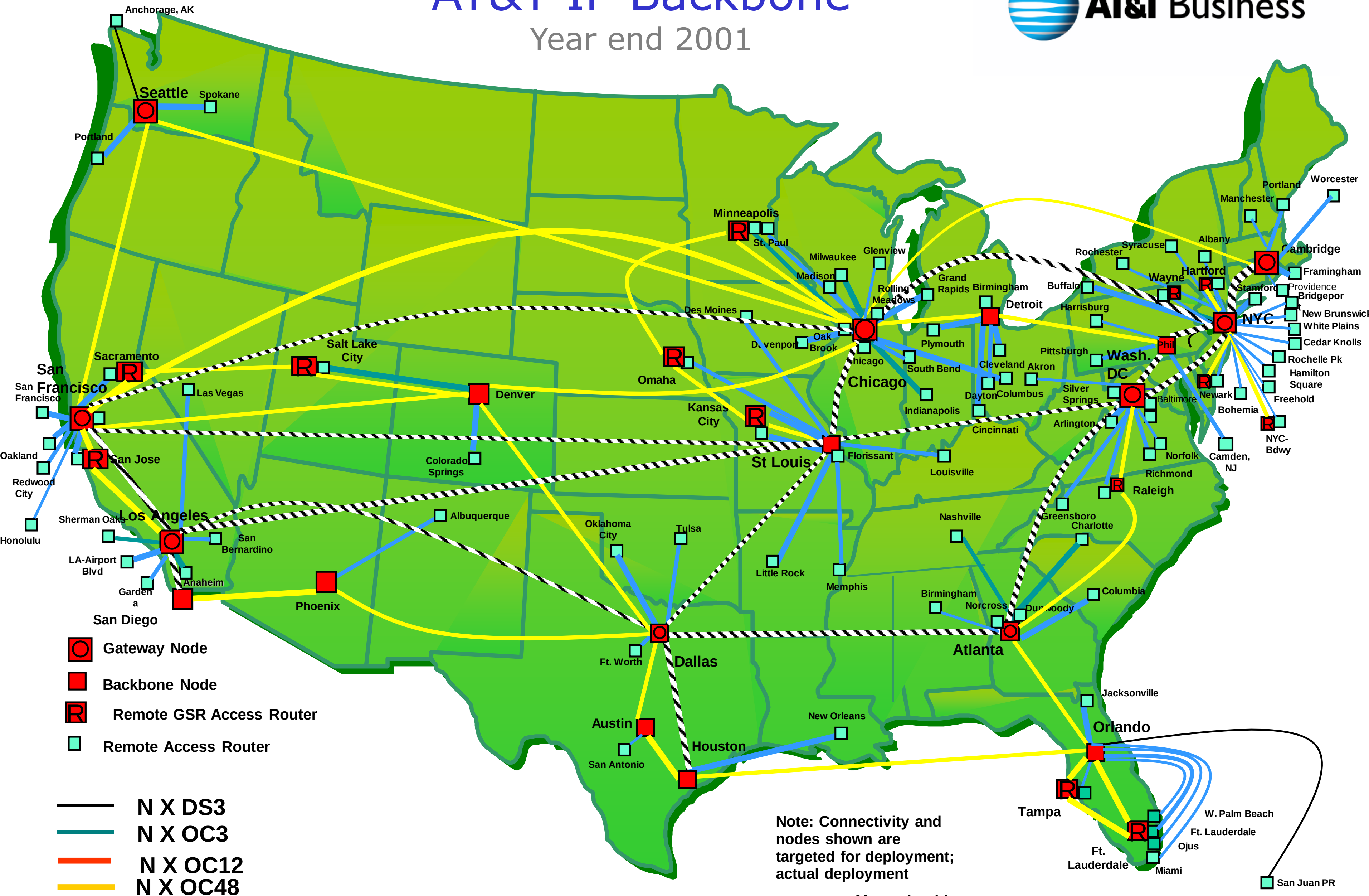


U.S. Sprint IP Backbone Network and Internet Centers (Q3 2001)
Seattle
Tacoma
To Pearl City, HI
Sacramento
Stockton
San Jose
Cheyenne
Chicago
Kansas City
Anaheim
Fort Worth
Dallas
Roachdale
Atlanta
Orlando
Springfield
Boston
New York
Pennsauken
Relay/DC
Reston
Raleigh
OC48
OC12
OC3
Internet Transport Node
Internet Center
Sprint.
Sprint E|Solutions

AT&T IP Backbone

Year end 2001

AT&T Business

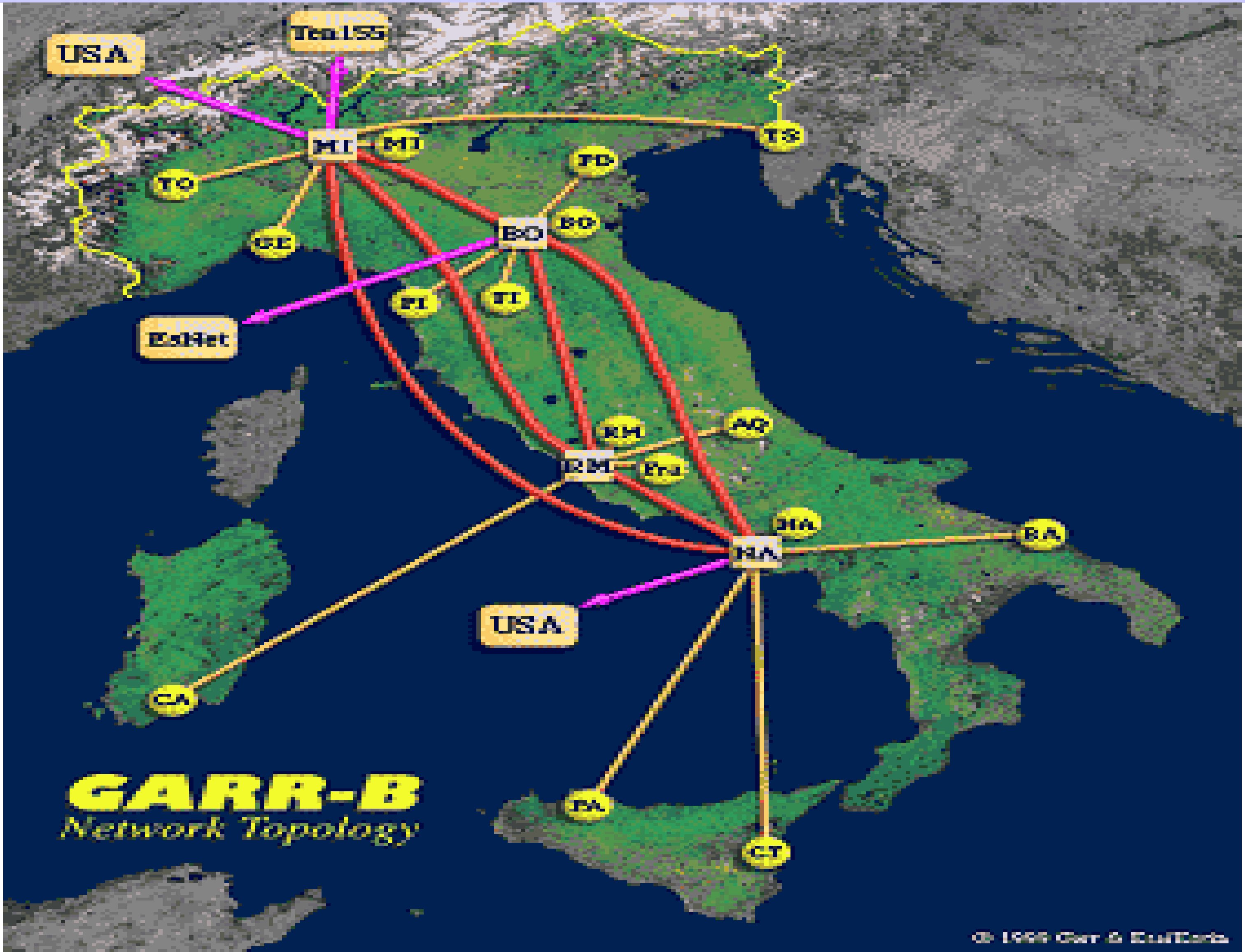


Note: Connectivity and nodes shown are targeted for deployment; actual deployment may vary. Maps should not be used to predict service availability.

ICNP 2002

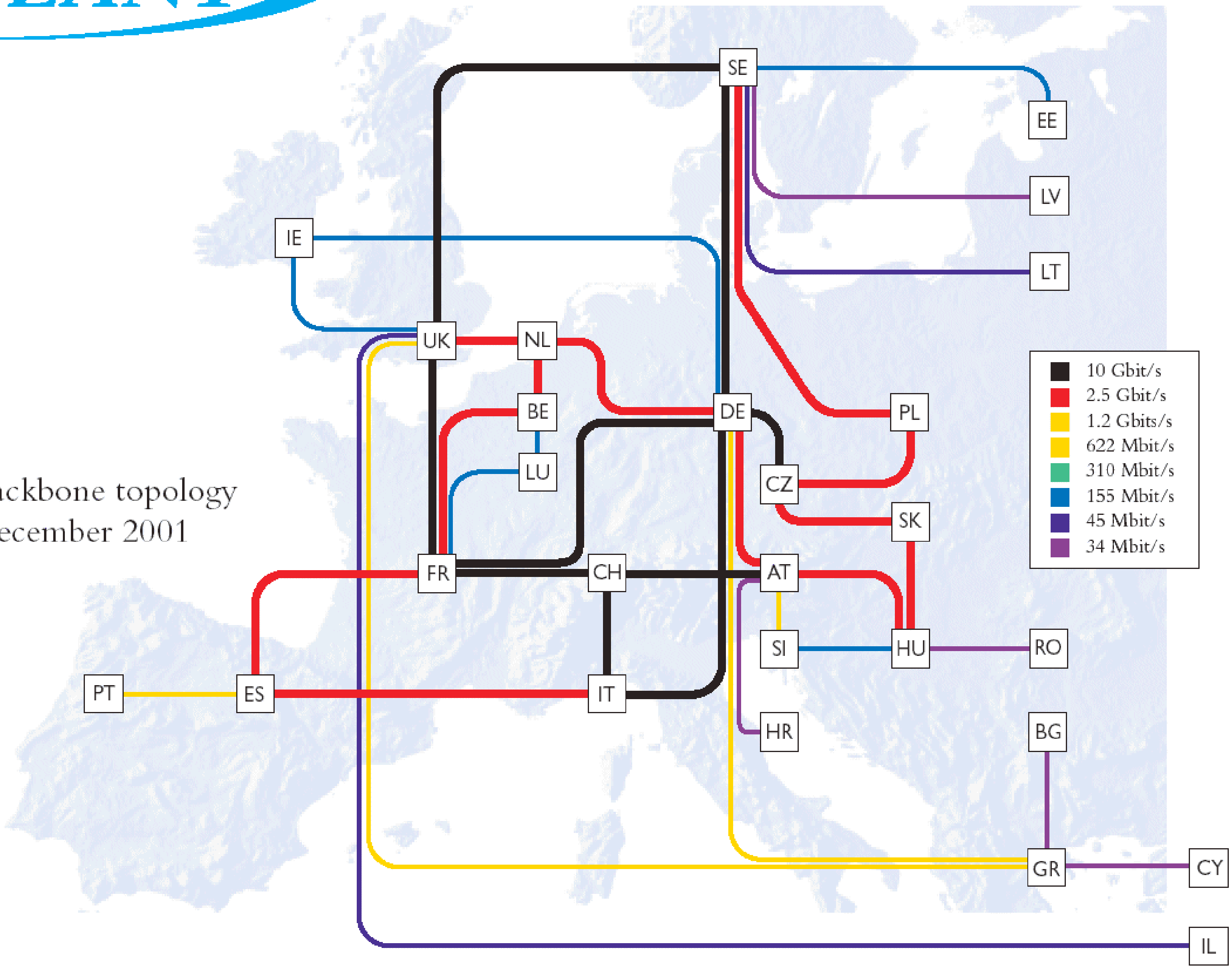
Rev. 6-4-01

GARR-B



USA
Ten155
MI
MI
TS
PD
TO
GE
BO
BO
FI
PI
EsNet
RM
AQ
RM
NA
BA
USA
CA
PA
CT
GARR-B
Network Topology

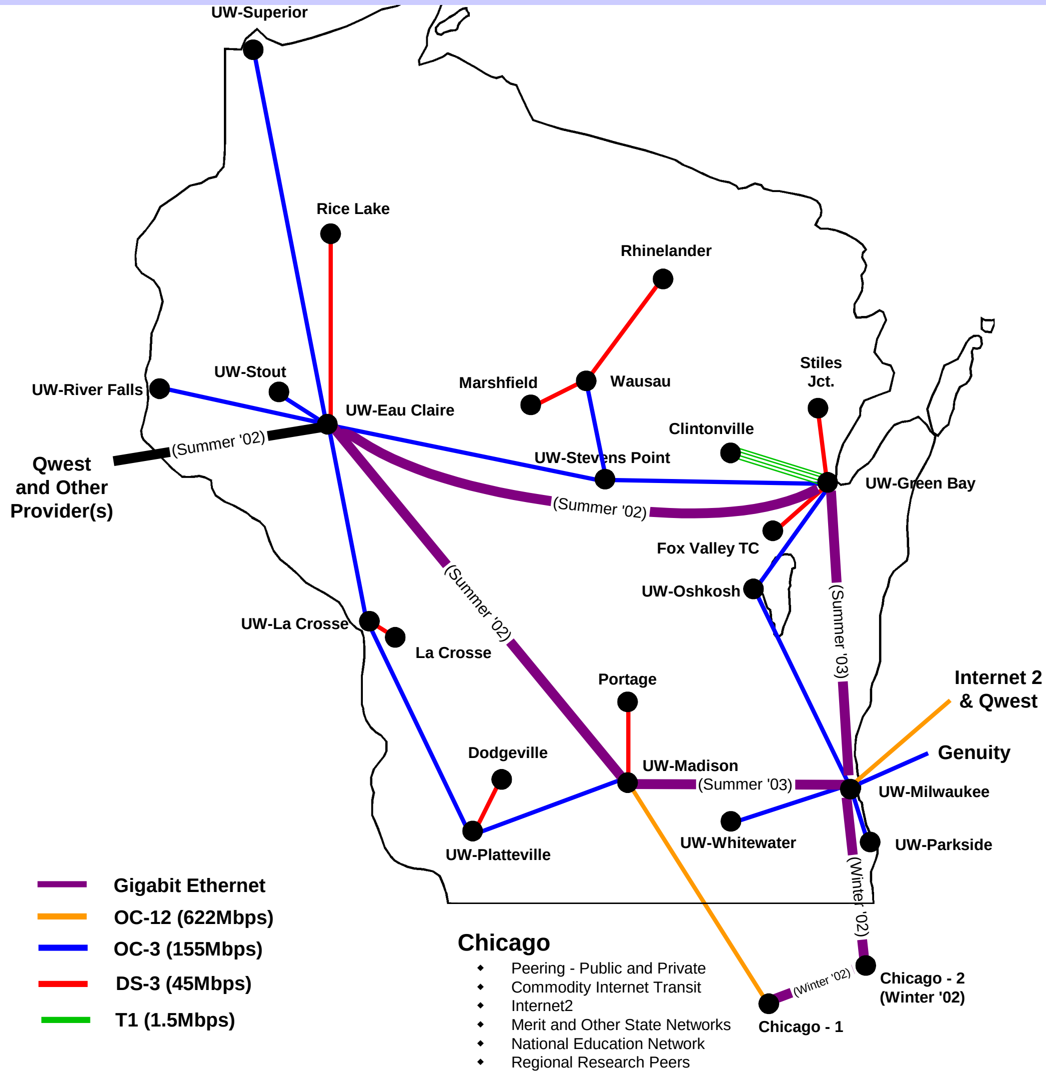
The Gigabit Research Network



Backbone topology
December 2001

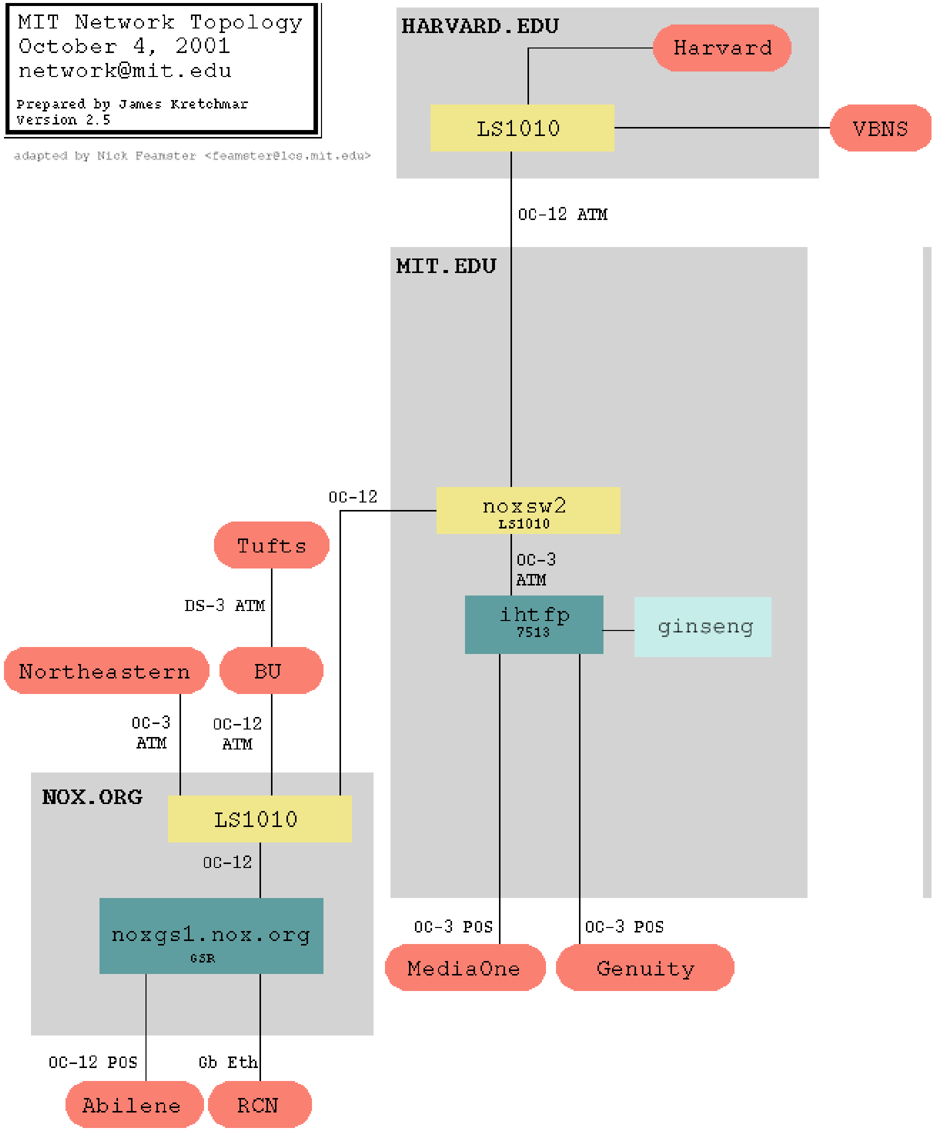
wiscnet.net

GO BUCKY!



ICNP 2002

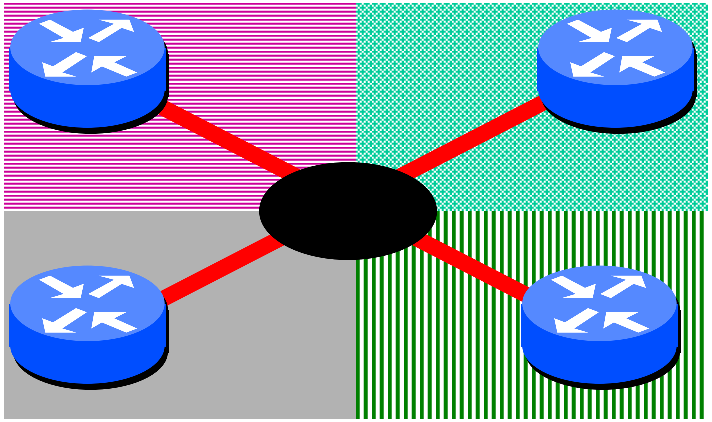
MIT.edu



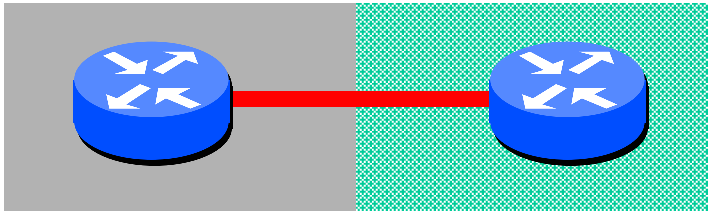
http://bgp.lcs.mit.edu/

Network Interconnections

- Exchange Point
  - Layer 2 or Layer 3



- Private Circuit
  - May be provided by a third party



PART II

Relationships Between Networks

Some Costs of Running an ISP

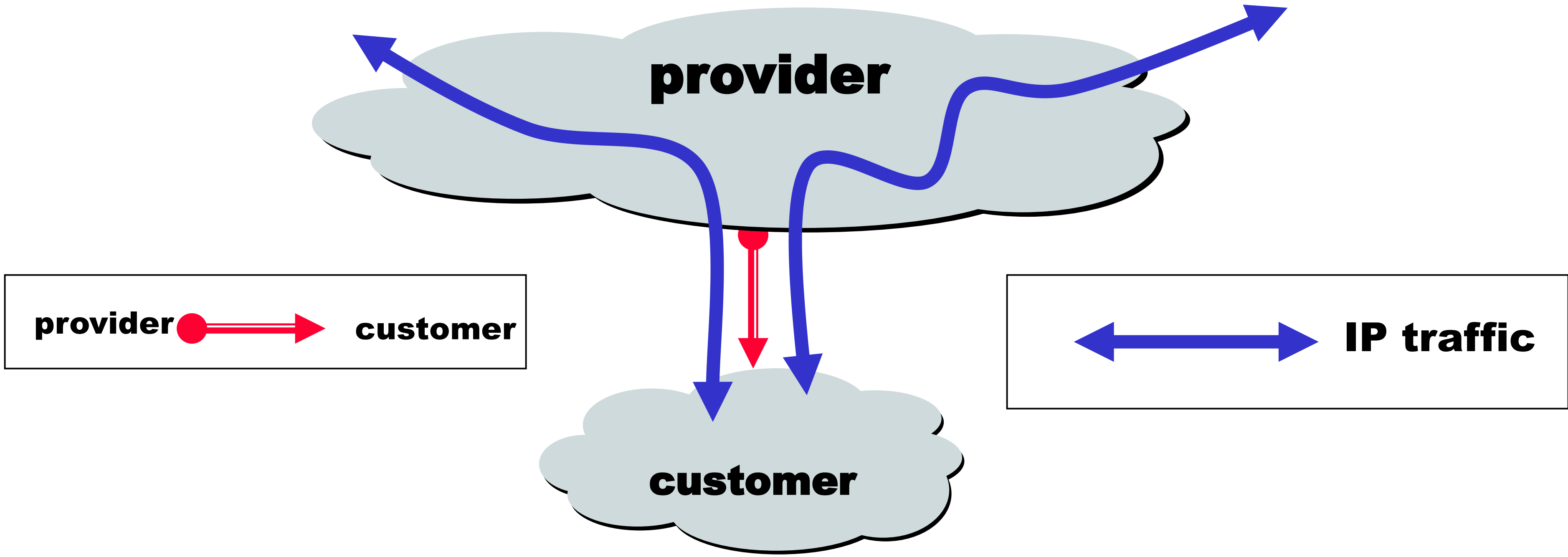
- People
- Physical connectivity and bandwidth
- Hardware
- Data center space and power
- …

Ballpark Figures (In US $)

- Hardware for an OC192 Pop: about 3,000,000.
  - Installation: 10,000
  - Power: 20,000/month
- OC192 link from NYC to D.C.: about 2,000,000/year
- Gigabit Ethernet IP connectivity
  - For end user: 10,000/month
  - For ISP: 30,000/month

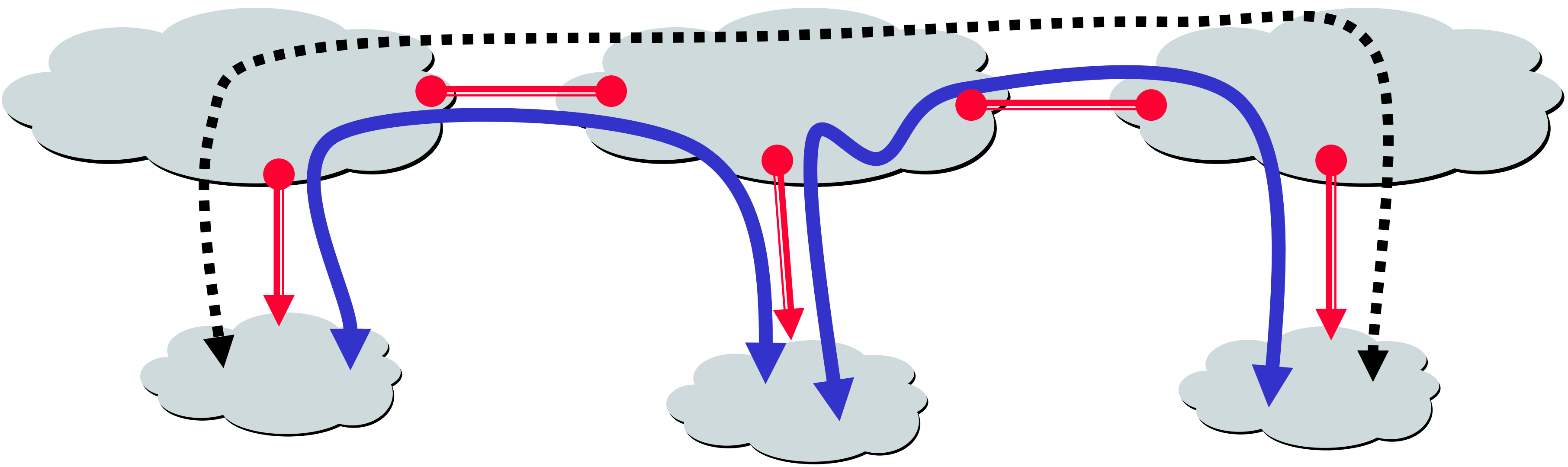
Prices can vary widely. Thanks to Ben Black and Vijay Gill for hints.

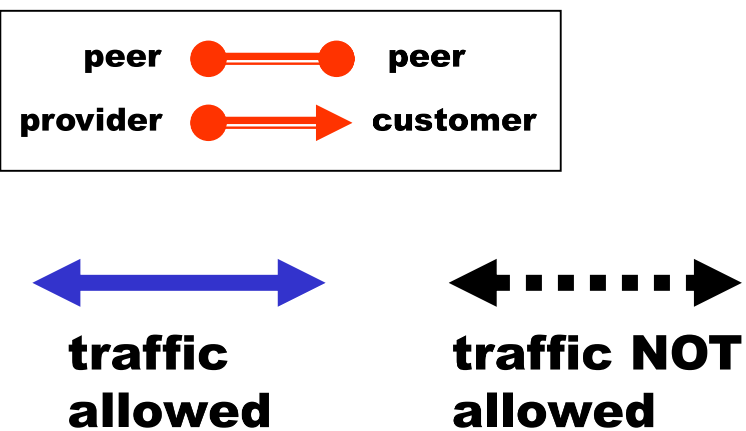
Customers and Providers



Customer pays provider for access to the Internet

The “Peering” Relationship



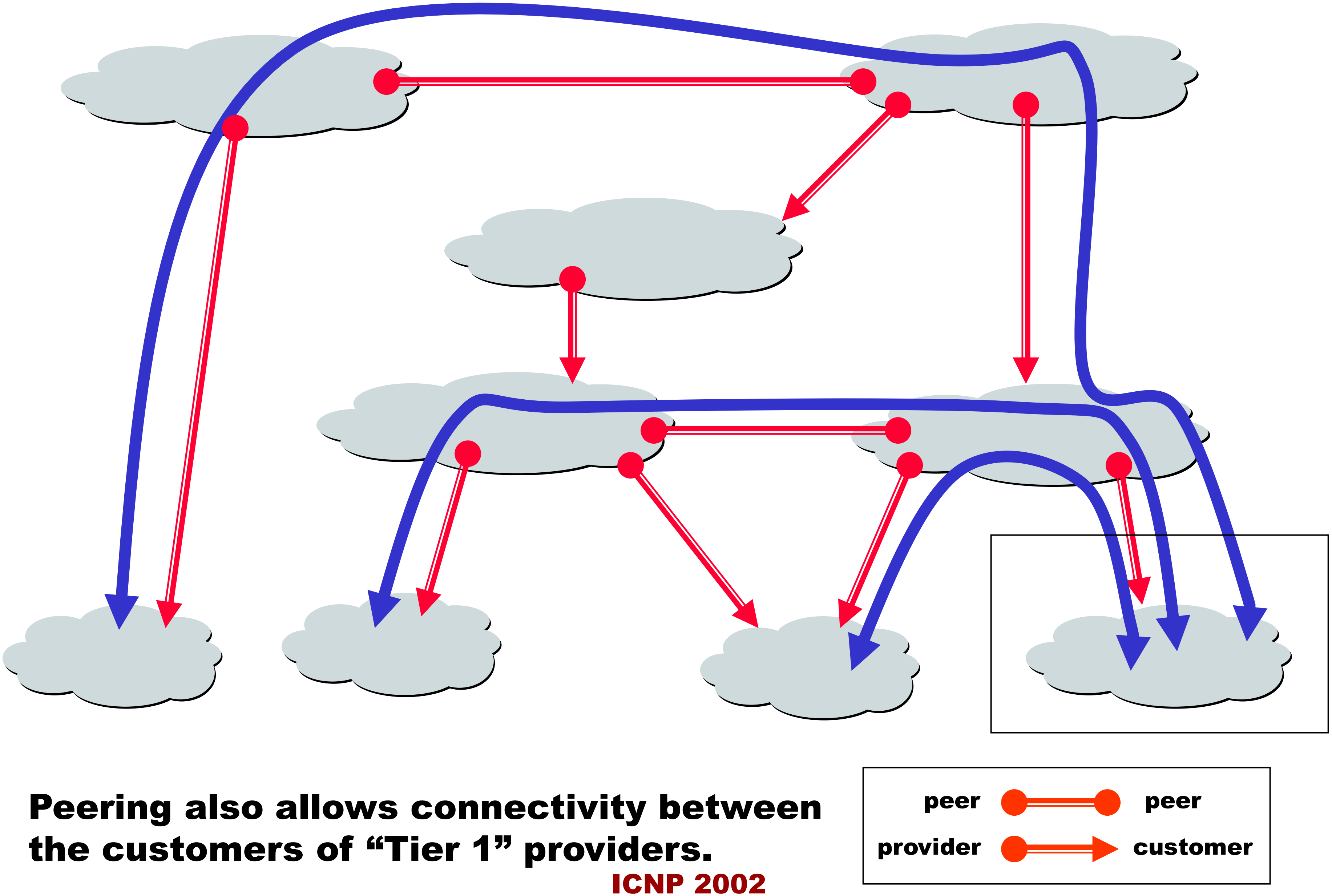


Peers provide transit between their respective customers

<u>Peers do not provide transit between peers</u>

Peers (often) do not exchange $$$

Peering Provides Shortcuts



Peering also allows connectivity between the customers of “Tier 1” providers.

Peering Wars

Peer

- Reduces upstream transit costs
- Can increase end-to-end performance
- May be the only way to connect your customers to some part of the Internet ("Tier 1")

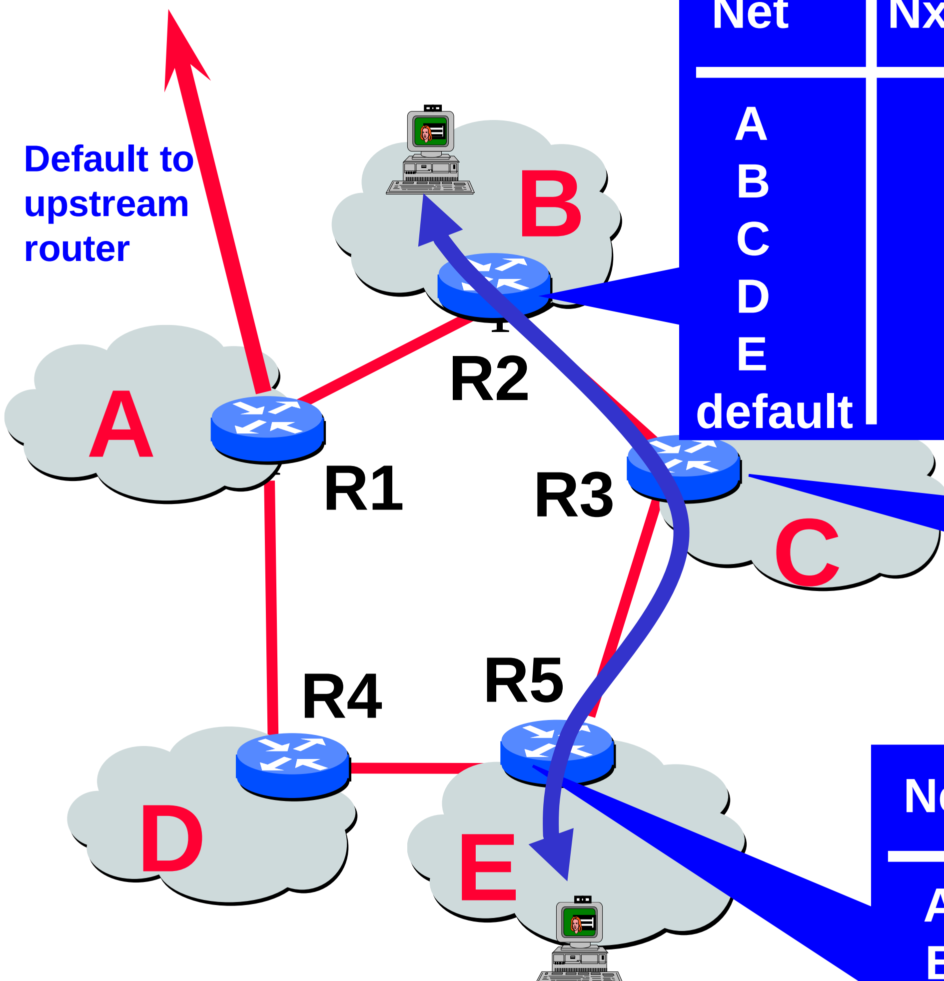
Don't Peer

- You would rather have customers
- Peers are usually your competition
- Peering relationships may require periodic renegotiation

Peering struggles are by far the most contentious issues in the ISP world!

Peering agreements are often confidential.

Routing vs. Forwarding



Net	Nxt Hop
A	R1
B	Direct
C	R3
D	R1
E	R3
default	R1

Forwarding always works

Routing can be badly broken

Net	Nxt Hop
A	R2
B	R2
C	Direct
D	R5
E	R5
default	R2

Net	Nxt Hop
A	R4
B	R3
C	R3
D	R4
E	Direct
default	R4

Forwarding: determine next hop

Routing: establish end-to-end paths

How Are Forwarding Tables Populated to implement Routing?

Statically

Administrator manually configures forwarding table entries

+ More control
+ Not restricted to destination-based forwarding
- Doesn't scale
- Slow to adapt to network failures

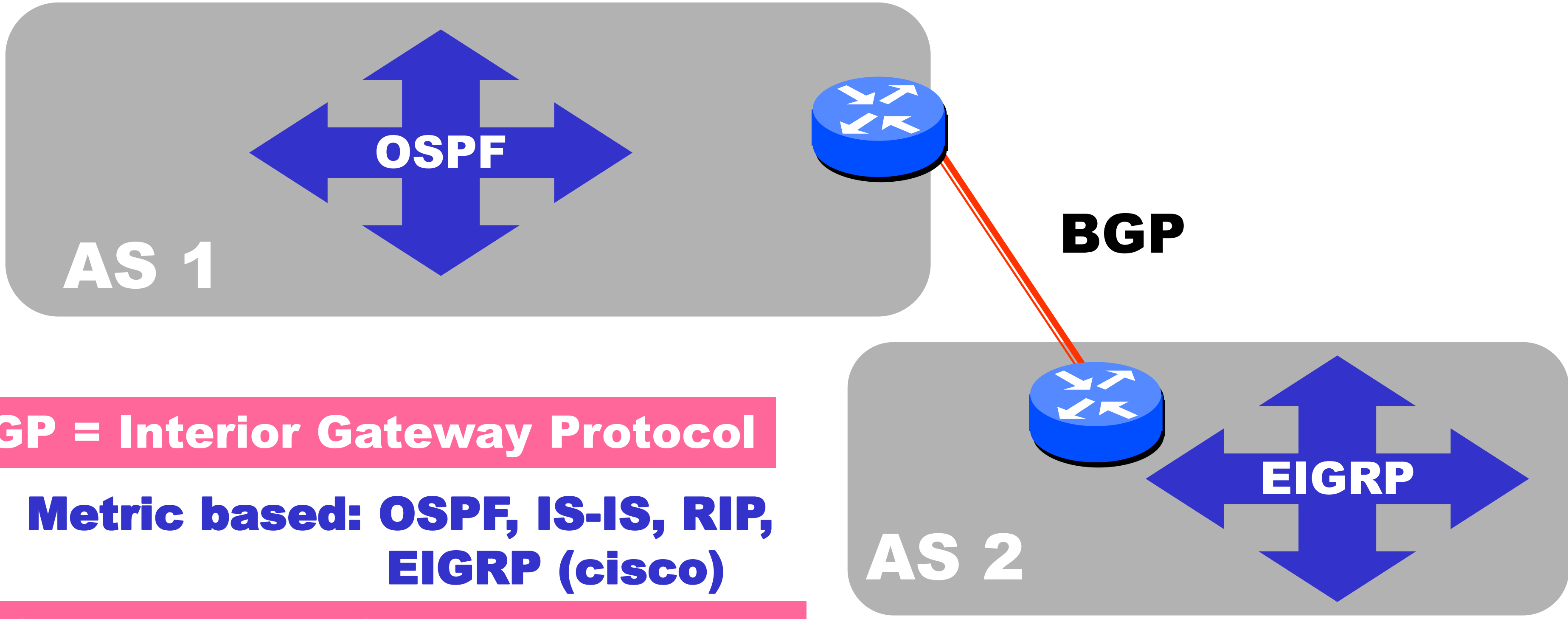
Dynamically

Routers exchange network reachability information using ROUTING PROTOCOLS. Routers use this to compute best routes

+ Can rapidly adapt to changes in network topology
+ Can be made to scale well
- Complex distributed algorithms
- Consume CPU, Bandwidth, Memory
- Debugging can be difficult
- Current protocols are destination-based

In practice : a mix of these.
Static routing mostly at the "edge"

Architecture of Dynamic Routing



IGP = Interior Gateway Protocol

Metric based: OSPF, IS-IS, RIP, EIGRP (cisco)

EGP = Exterior Gateway Protocol

Policy based: BGP

The Routing Domain of BGP is the entire Internet

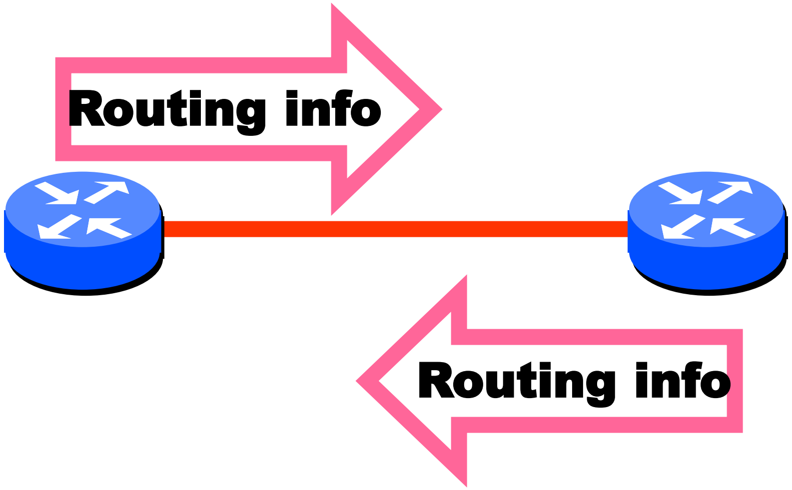
Technology of Distributed Routing

Link State	Vectoring
• Topology information is flooded within the routing domain	• Each router knows little about network topology
• Best end-to-end paths are computed locally at each router.	• Only best next-hops are chosen by each router for each destination network.
• Best end-to-end paths determine next-hops.	• Best end-to-end paths result from composition of all next-hop choices
• Based on minimizing some notion of distance	• Does not require any notion of distance
• Works only if policy is shared and uniform	• Does not require uniform policies at all routers
• Examples: OSPF, IS-IS	• Examples: RIP, BGP

The Gang of Four

	Link State	Vectoring
IGP	OSPF IS-IS	RIP
EGP		BGP

Routers Talking to Routers



- Routing computation is distributed among routers within a routing domain
- Computation of best next hop based on routing information is the most CPU/memory intensive task on a router
- Routing messages are usually not routed, but exchanged via layer 2 between physically adjacent routers (internal BGP and multi-hop external BGP are exceptions)

Autonomous Routing Domains (ARDs)

A collection of physical networks glued together using IP, that have a unified administrative routing policy.

- **Campus networks**
- **Corporate networks**
- **ISP Internal networks**
- **…**

Autonomous Systems (ASes)

An autonomous system is an autonomous routing domain that has been assigned an Autonomous System Number (ASN).

… the administration of an AS appears to other ASes to have a single coherent interior routing plan and presents a consistent picture of what networks are reachable through it.

RFC 1930: Guidelines for creation, selection, and registration of an Autonomous System

AS Numbers (ASNs)

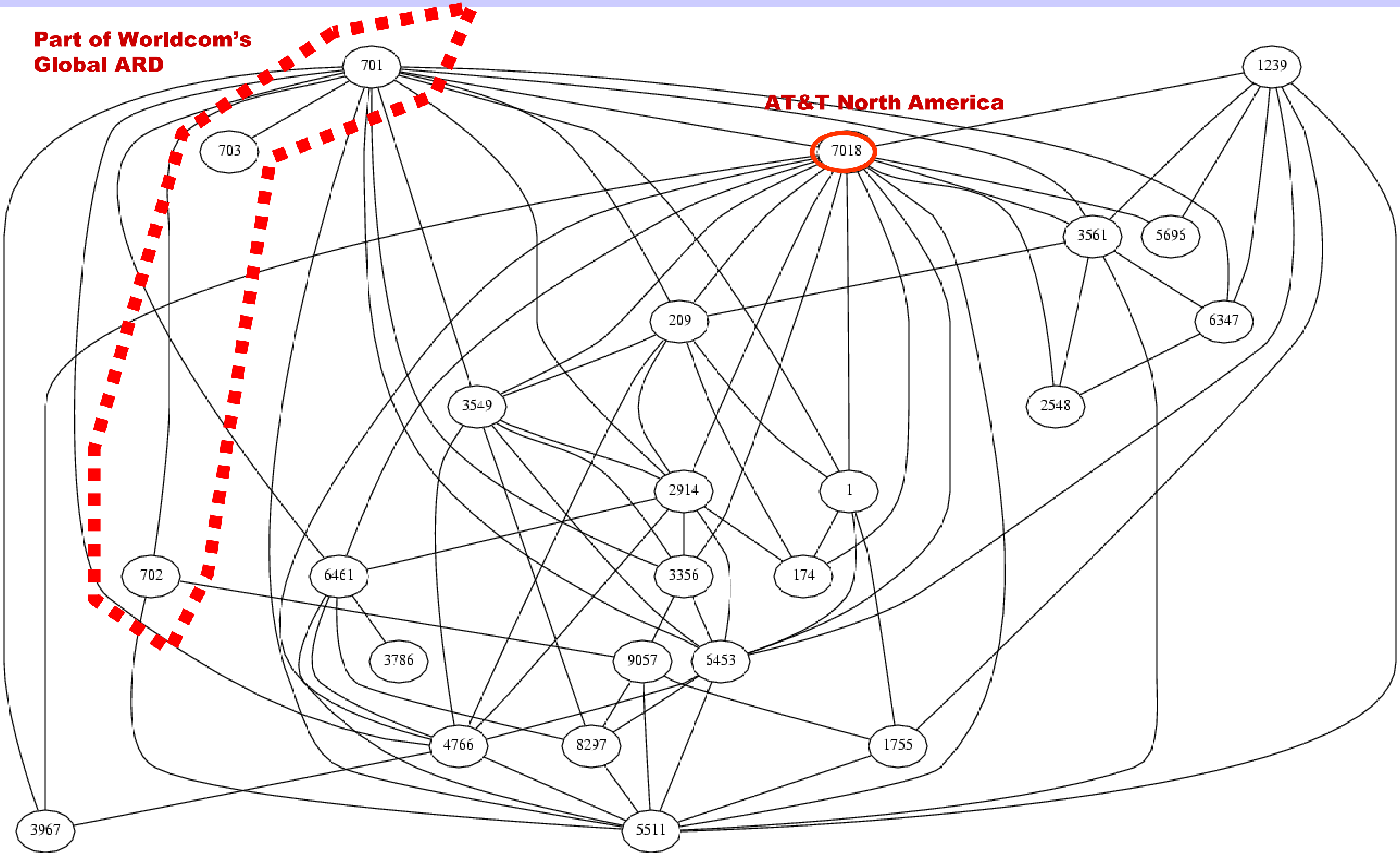
ASNs are 16 bit values.
64512 through 65535 are “private”

Currently over 11,000 in use.

- **Genuity (f.k.a. BBN): 1**
- **MIT: 3**
- **Harvard: 11**
- **UC San Diego: 7377**
- **AT&T: 7018, 6341, 5074, …**
- **UUNET: 701, 702, 284, 12199, …**
- **Sprint: 1239, 1240, 6211, 6242, …**
- **…**

ASNs represent units of routing policy

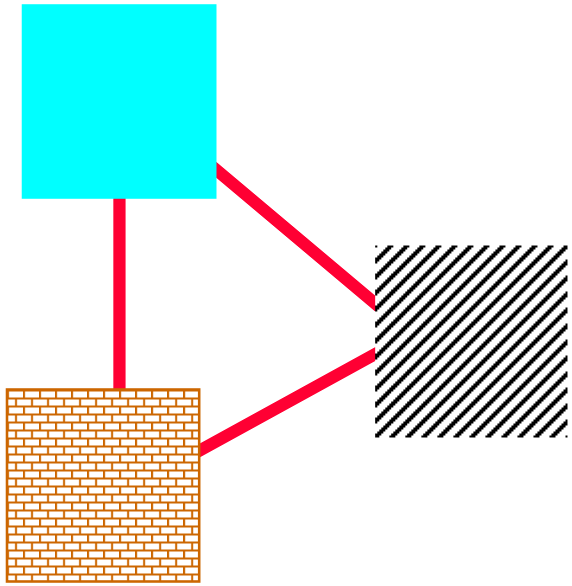
AS Graphs Can Be Fun

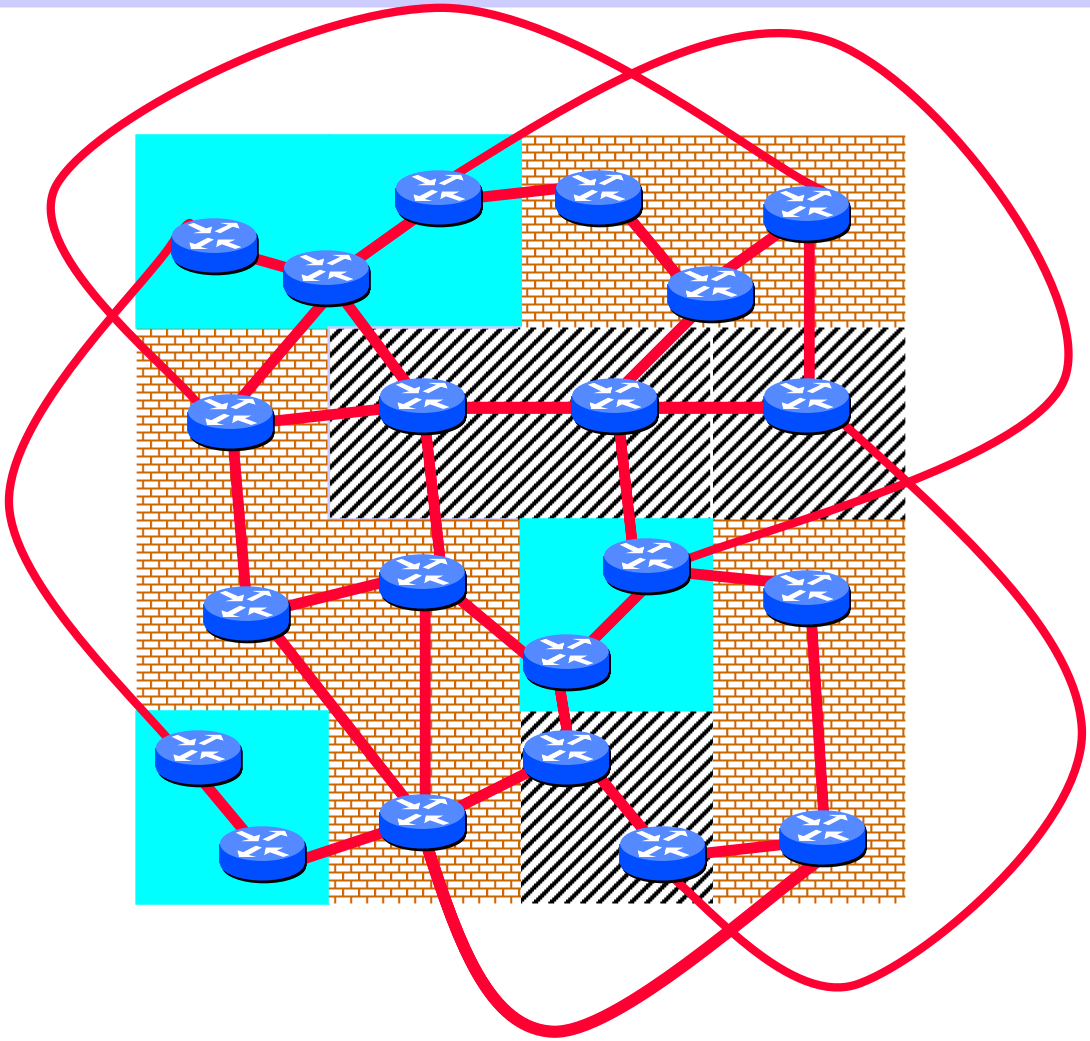


The subgraph showing all ASes that have more than 100 neighbors in full graph of 11,158 nodes. July 6, 2001. Point of view: AT&T route-server

AS Graph != Internet Topology

BGP was designed to throw away information!

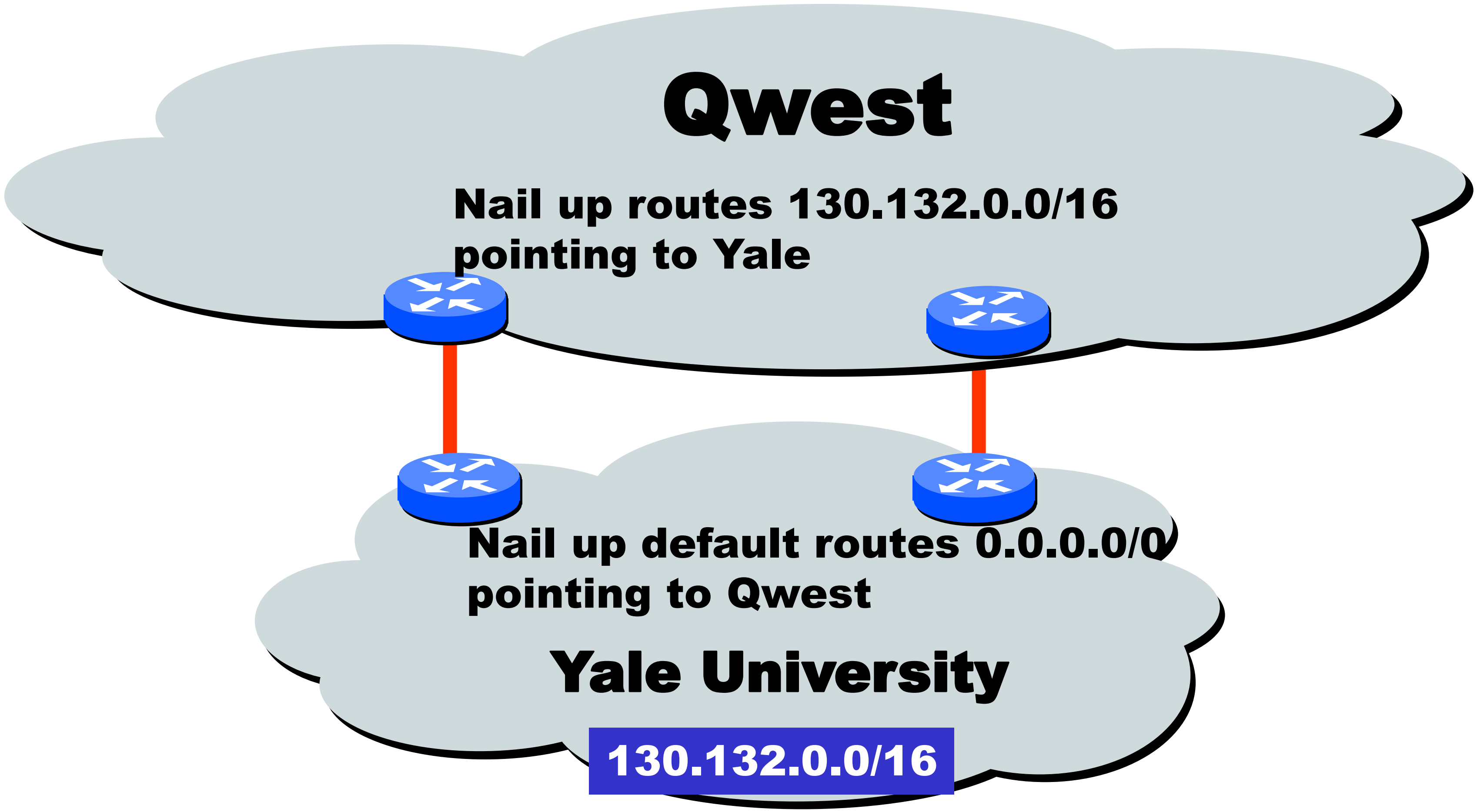




The AS graph may look like this.

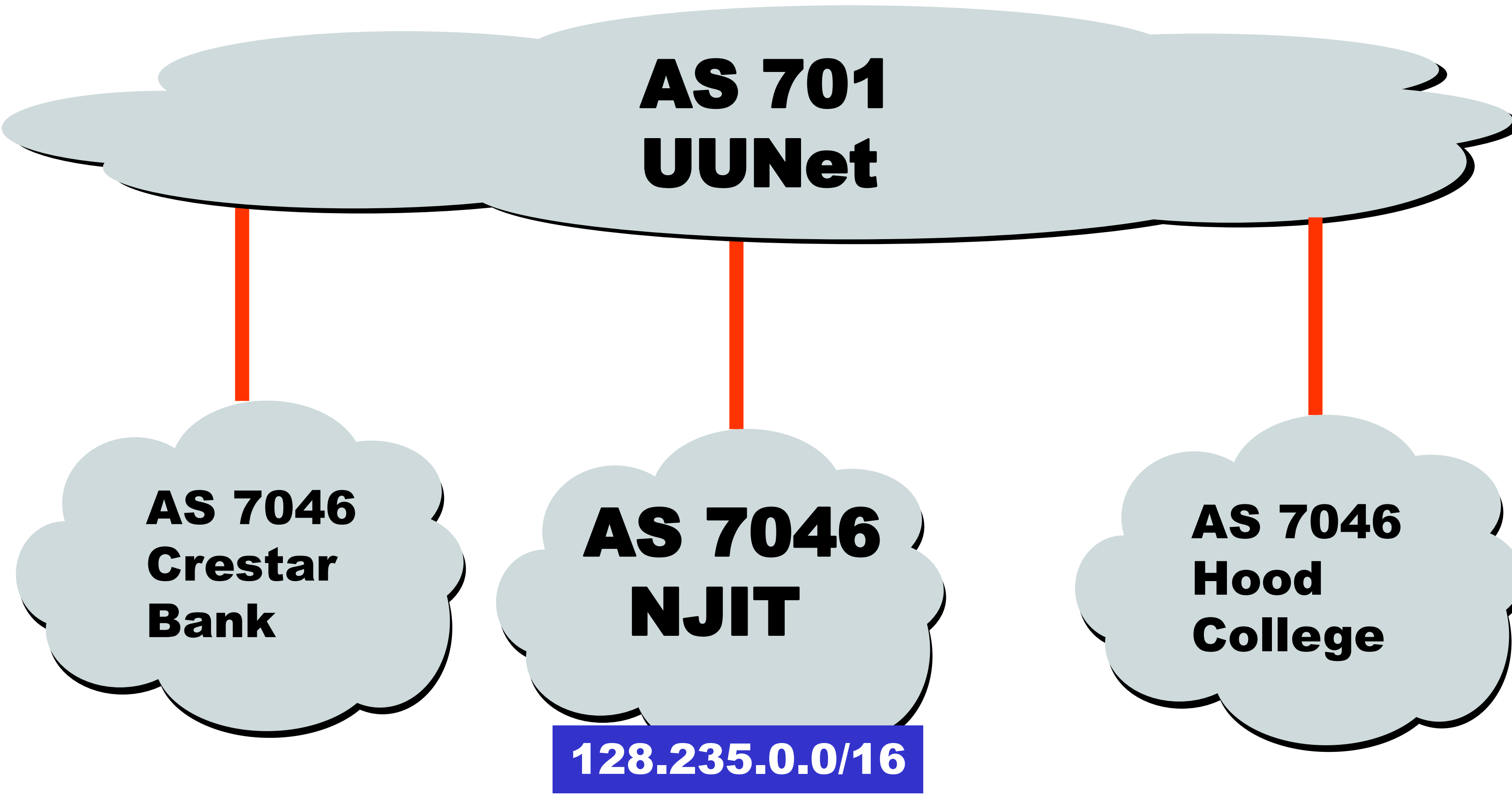
Reality may be closer to this…

Autonomous Routing Domains Don't Always Need BGP or an ASN



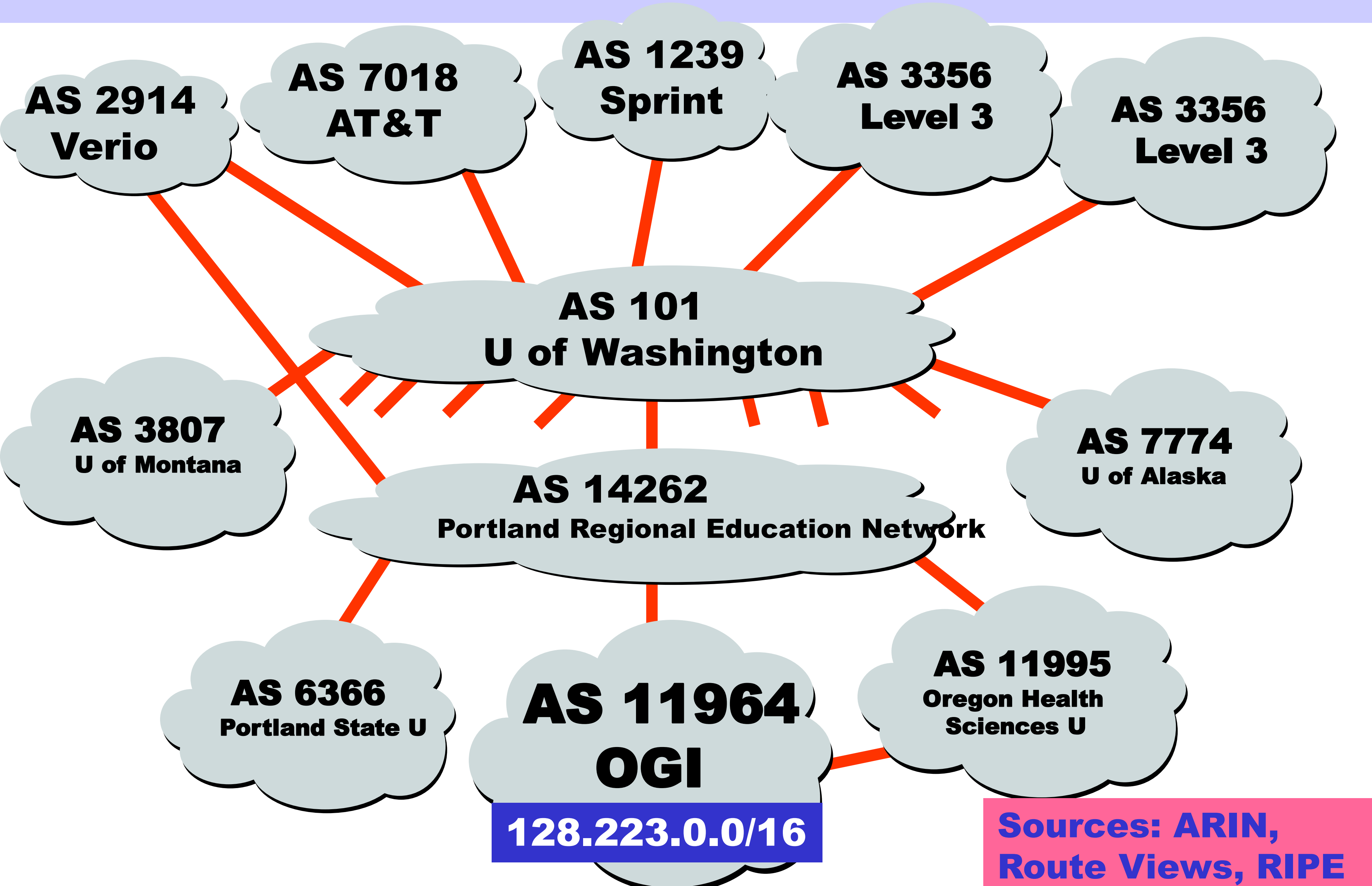
Static routing is the most common way of connecting an autonomous routing domain to the Internet.
This helps explain why BGP is a mystery to many …

ASNs Can Be “Shared” (RFC 2270)



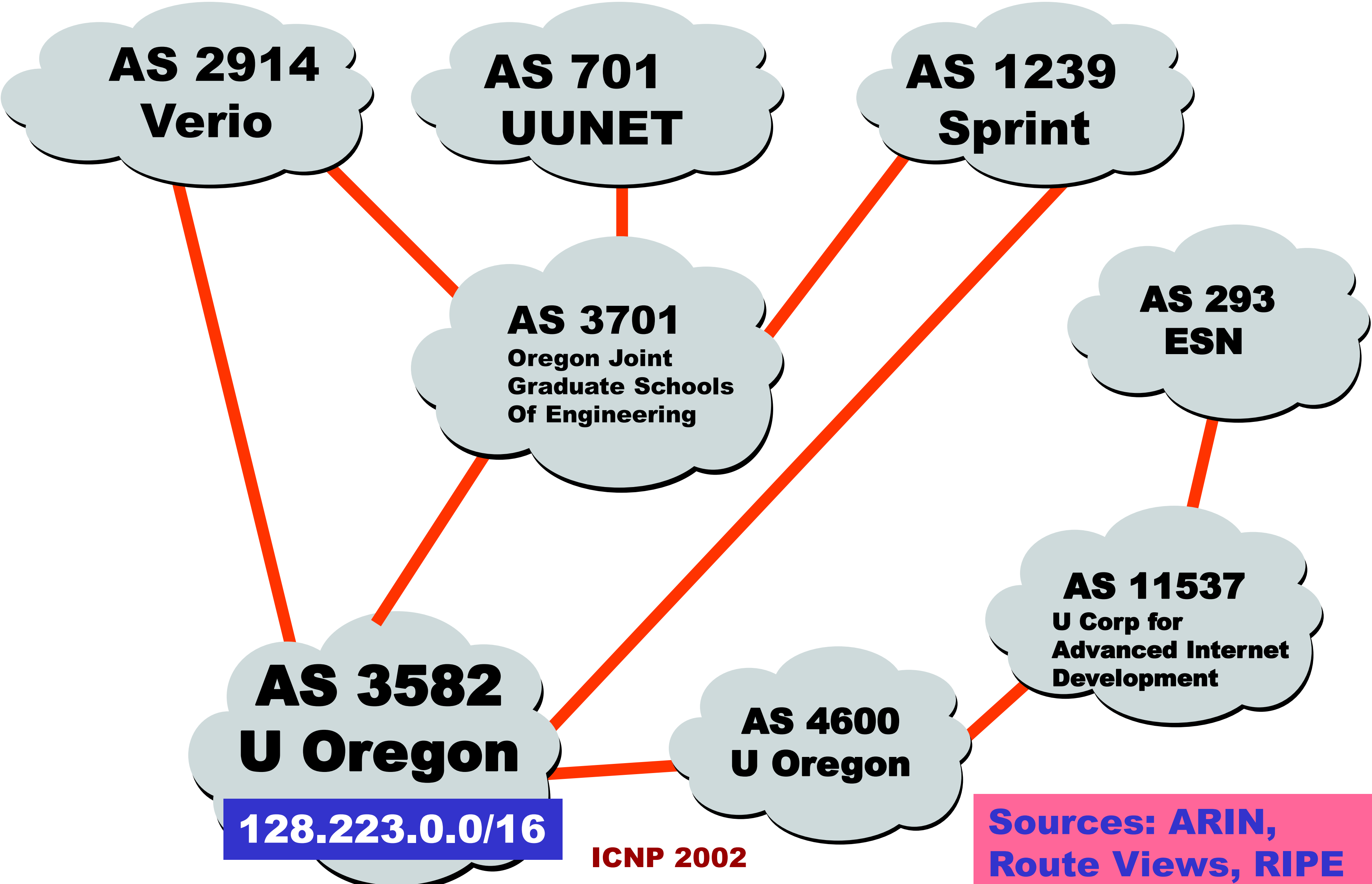
ASN 7046 is assigned to UUNet. It is used by Customers single homed to UUNet, but needing BGP for some reason (load balancing, etc..) [RFC 2270]

A Bit of OGI's AS Neighborhood



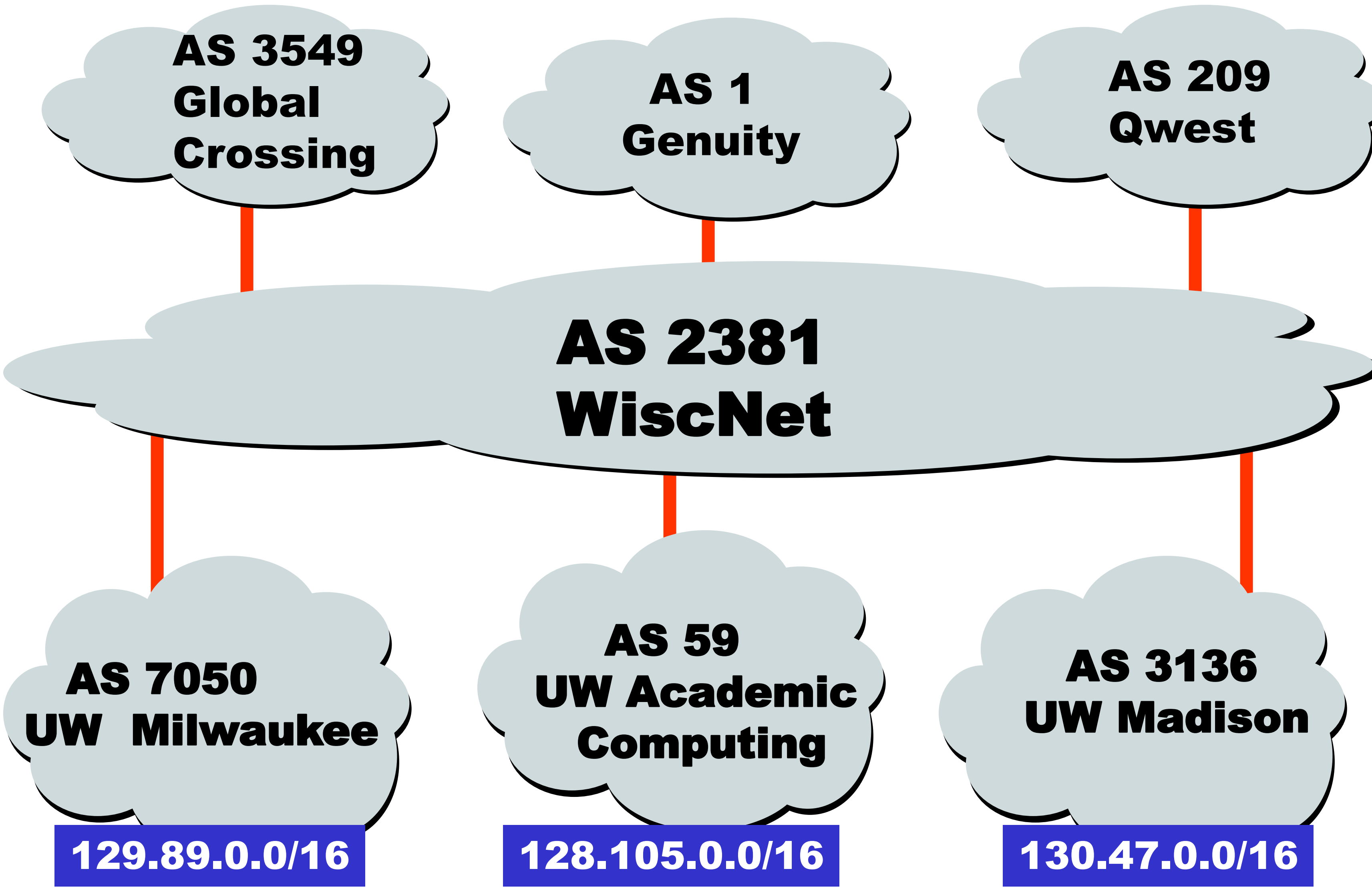
Sources: ARIN, Route Views, RIPE

A Bit of U Oregon's AS Neighborhood



Sources: ARIN, Route Views, RIPE

Partial View of cs.wisc.edu Neighborhood



ARD != AS

- Most ARDs have no ASN (statically routed at Internet edge)
- Some unrelated ARDs share the same ASN (RFC 2270)
- Some ARDs are implemented with multiple ASNs (example: Worldcom)

ASes are an implementation detail of Interdomain routing

PART III

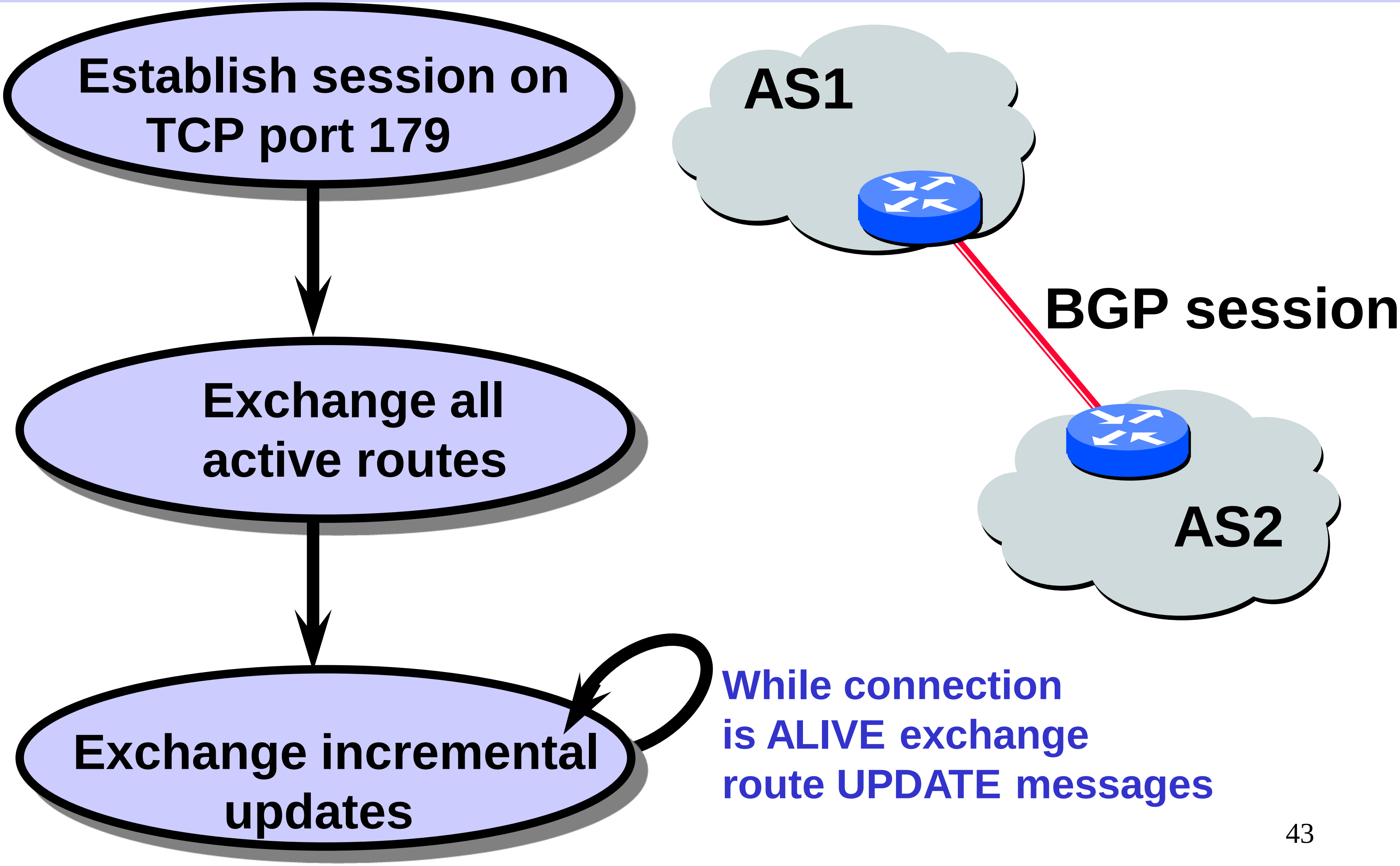
Implementing Inter-Network Relationships with BGP

BGP-4

- **BGP** = **<u>B</u>**order **<u>G</u>**ateway **<u>P</u>**rotocol
- Is a **<u>Policy-Based</u>** routing protocol
- Is the **<u>de facto EGP</u>** of today's global Internet
- Relatively simple protocol, but configuration is complex and the entire world can see, and be impacted by, your mistakes.

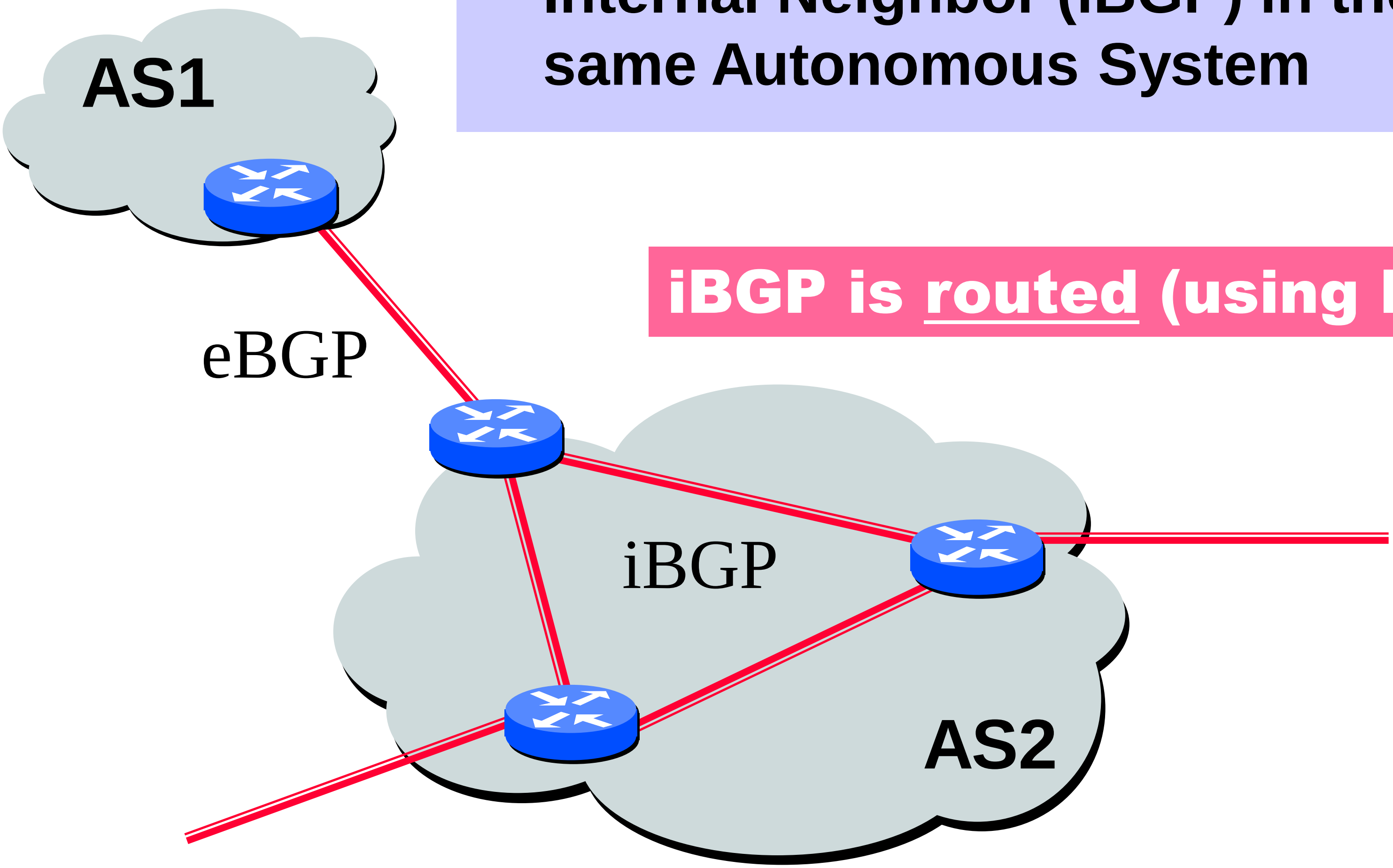
- **1989 : BGP-1 [RFC 1105]**
  - **Replacement for EGP (1984, RFC 904)**
- **1990 : BGP-2 [RFC 1163]**
- **1991 : BGP-3 [RFC 1267]**
- **1995 : BGP-4 [RFC 1771]**
  - **Support for Classless Interdomain Routing (CIDR)**

BGP Operations (Simplified)



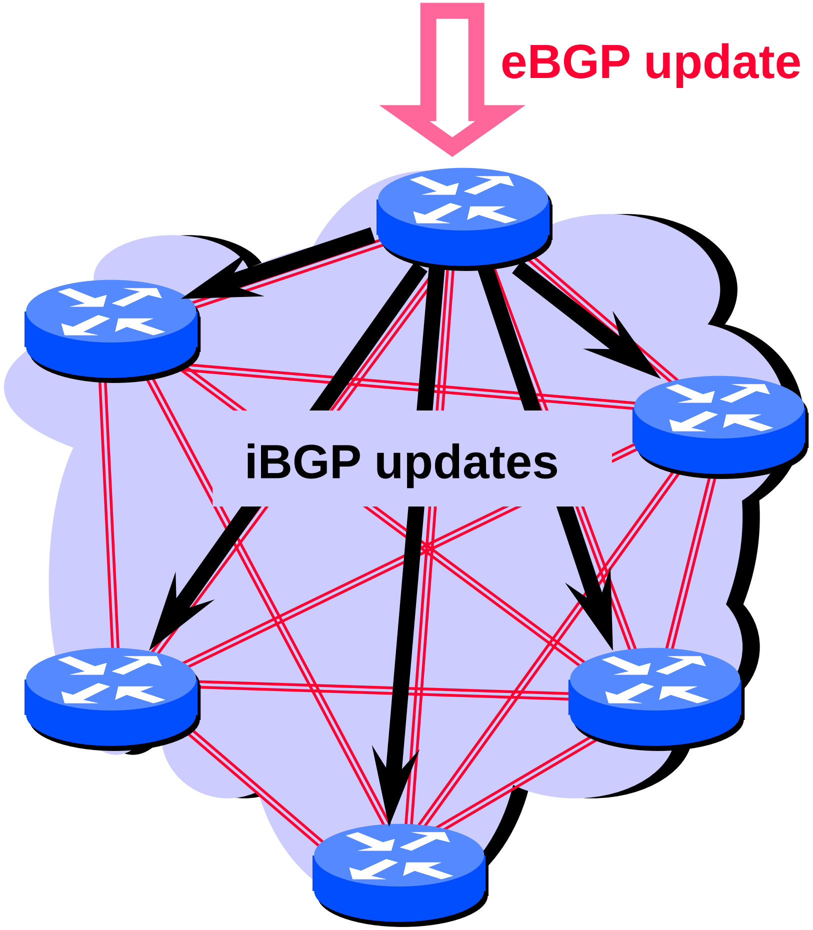
Two Types of BGP Neighbor Relationships

- **External Neighbor (eBGP) in a different Autonomous Systems**
- **Internal Neighbor (iBGP) in the same Autonomous System**



iBGP is <u>routed</u> (using IGP!)

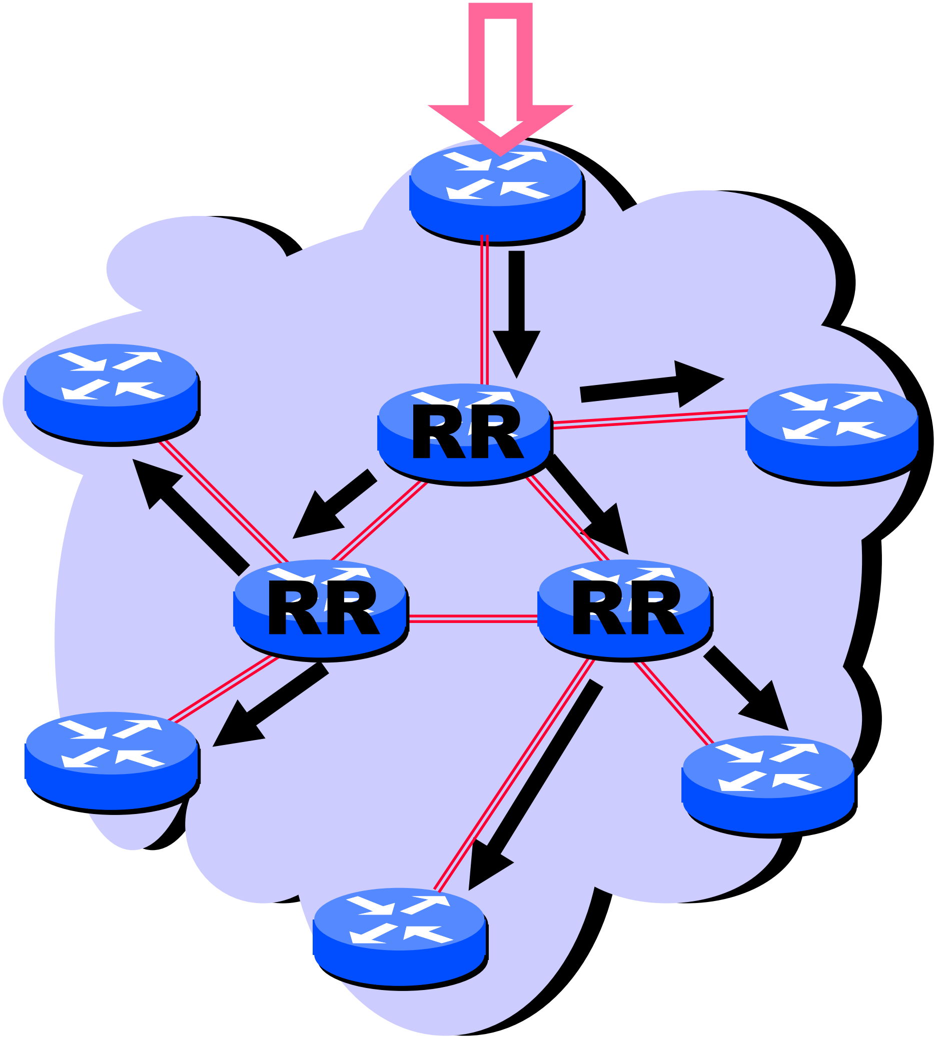
iBGP Mesh Does Not Scale



- N border routers means N(N-1)/2 peering sessions
- Each router must have N-1 iBGP sessions configured
- The addition a single iBGP speaker requires configuration changes to all other iBGP speakers
- Size of iBGP routing table can be order N larger than number of best routes (remember alternate routes!)
- Each router has to listen to update noise from each neighbor

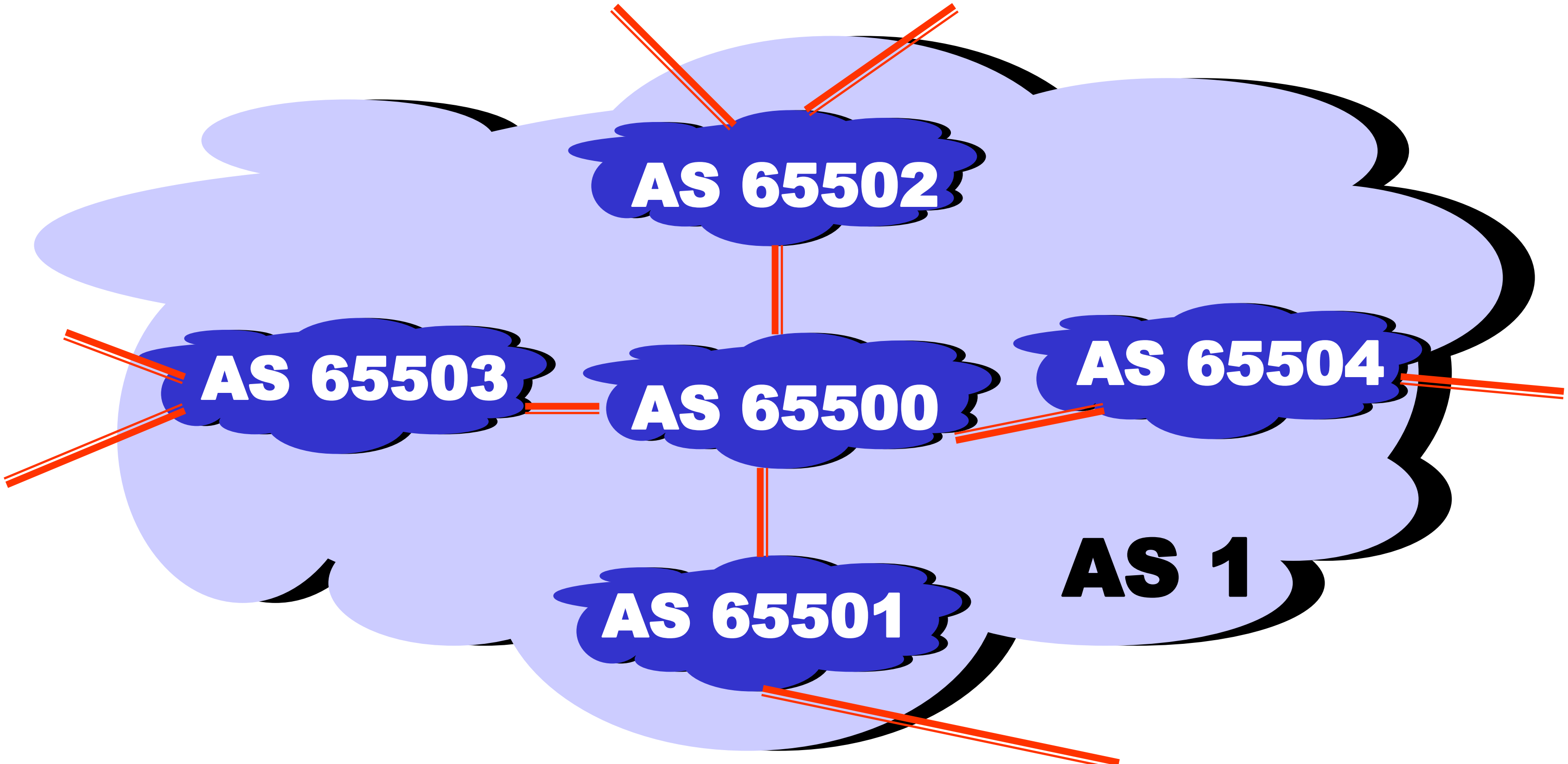
Currently four solutions:
(0) Buy bigger routers!
(1) Break AS into smaller ASes
(2) BGP Route reflectors
(3) BGP confederations

Route Reflectors



- Route reflectors can pass on iBGP updates to clients
- Each RR passes along ONLY best routes
- ORIGINATOR_ID and CLUSTER_LIST attributes are needed to avoid loops

BGP Confederations



From the outside, this looks like AS 1

Confederation eBGP (between member ASes) preserves LOCAL_PREF, MED, and BGP NEXTHOP.

Four Types of BGP Messages

- **Open : Establish a peering session.**
- **Keep Alive : Handshake at regular intervals.**
- **Notification : Shuts down a peering session.**
- **Update : Announcing new routes or withdrawing previously announced routes.**

announcement
=
prefix + attributes values

BGP Attributes

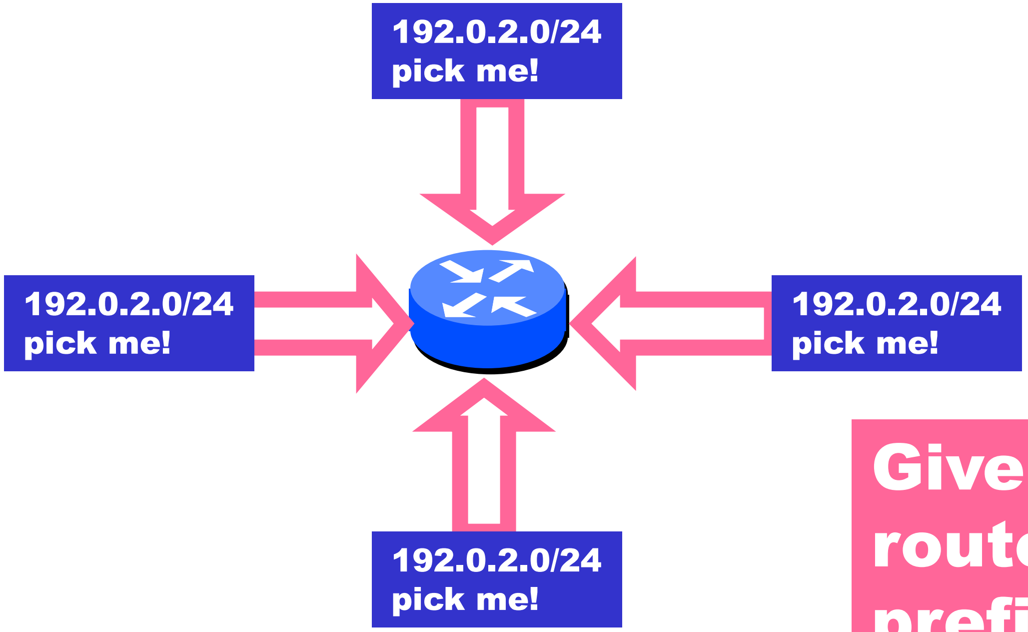
Value	Code	Reference
1	ORIGIN	[RFC1771]
2	AS_PATH	[RFC1771]
3	NEXT_HOP	[RFC1771]
4	MULTI_EXIT_DISC	[RFC1771]
5	LOCAL_PREF	[RFC1771]
6	ATOMIC_AGGREGATE	[RFC1771]
7	AGGREGATOR	[RFC1771]
8	COMMUNITY	[RFC1997]
9	ORIGINATOR_ID	[RFC2796]
10	CLUSTER_LIST	[RFC2796]
11	DPA	[Chen]
12	ADVERTISER	[RFC1863]
13	RCID_PATH / CLUSTER_ID	[RFC1863]
14	MP_REACH_NLRI	[RFC2283]
15	MP_UNREACH_NLRI	[RFC2283]
16	EXTENDED COMMUNITIES	[Rosen]
...		
255	reserved for development	

From IANA: http://www.iana.org/assignments/bgp-parameters

Most important attributes

Not all attributes need to be present in every announcement

Attributes are Used to Select Best Routes



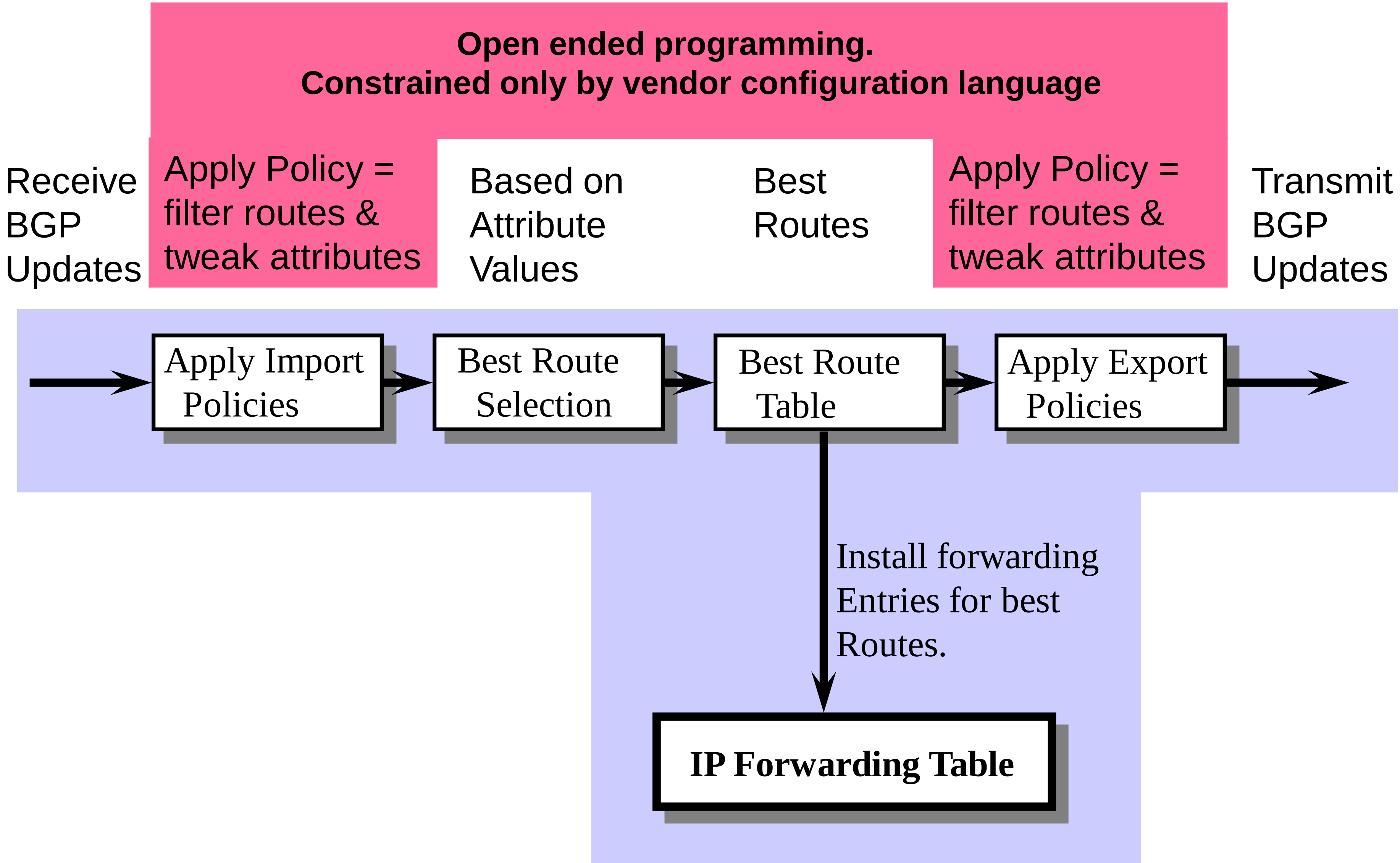
Given multiple routes to the same prefix, a BGP speaker must pick at most <u>one</u> best route

(Note: it could reject them all!)

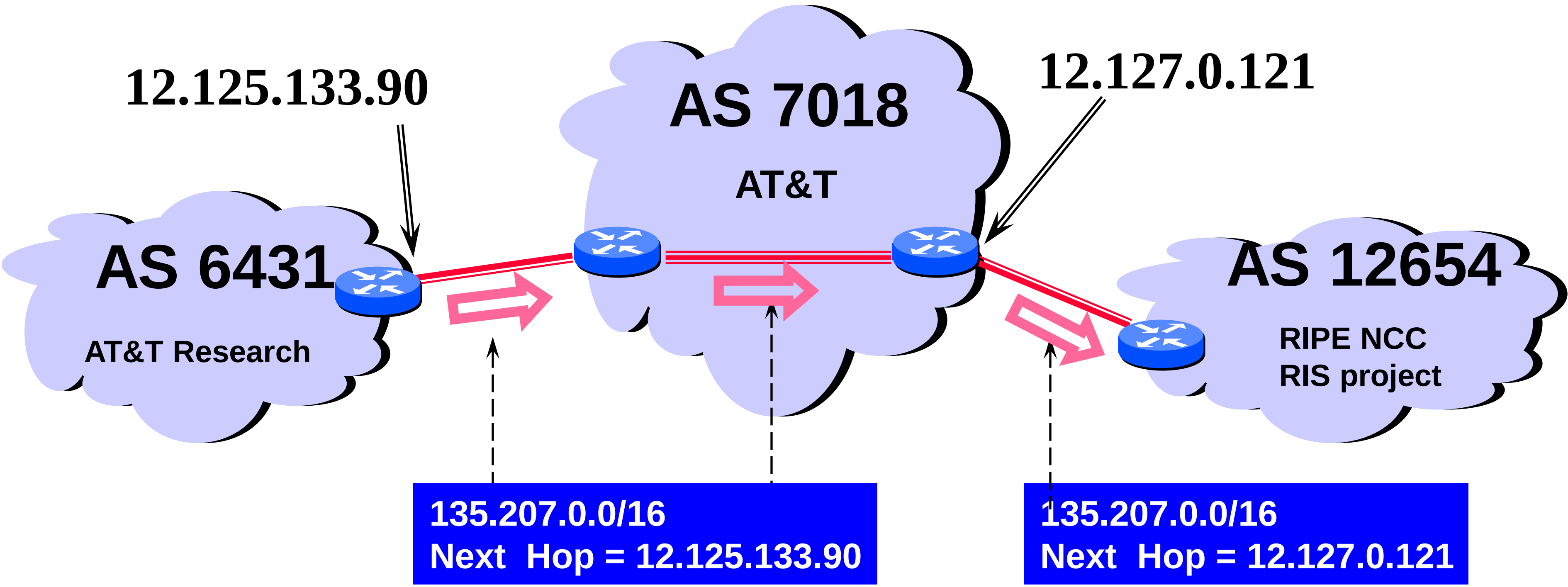
Route Selection Summary

Highest Local Preference	Enforce relationships
Shortest ASPATH **Lowest MED** **i-BGP < e-BGP** **Lowest IGP cost to BGP egress**	traffic engineering
Lowest router ID	Throw up hands and break ties

BGP Route Processing

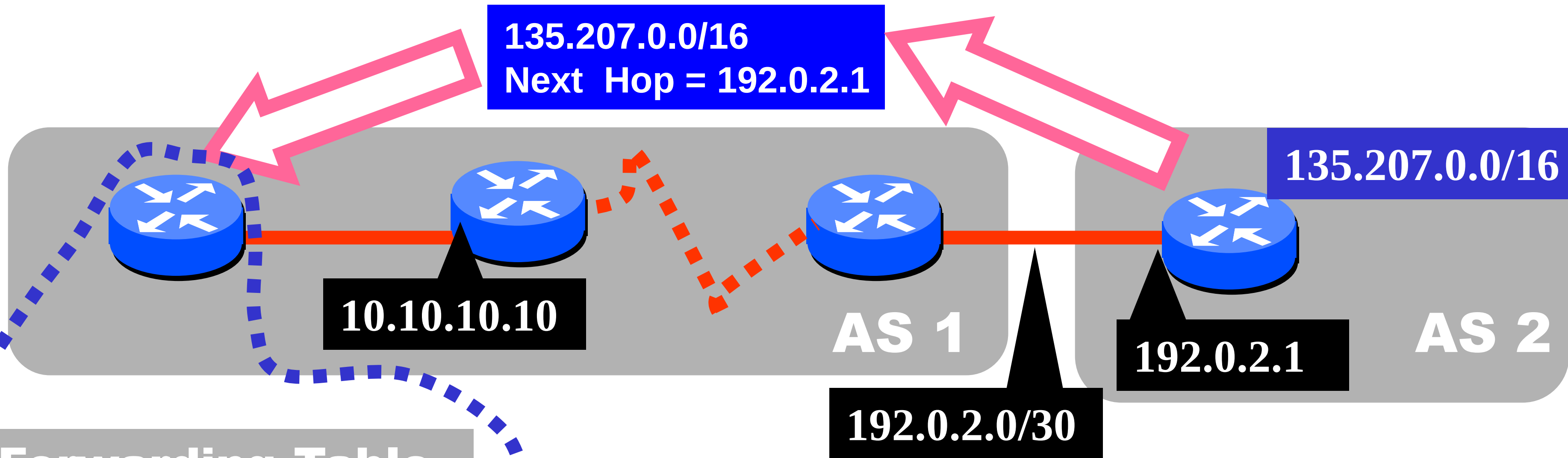


BGP Next Hop Attribute



Every time a route announcement crosses an AS boundary, the Next Hop attribute is changed to the IP address of the border router that announced the route.

Join EGP with IGP For Connectivity



Forwarding Table

destination	next hop
192.0.2.0/30	10.10.10.10

+

EGP

destination	next hop
135.207.0.0/16	192.0.2.1

Forwarding Table

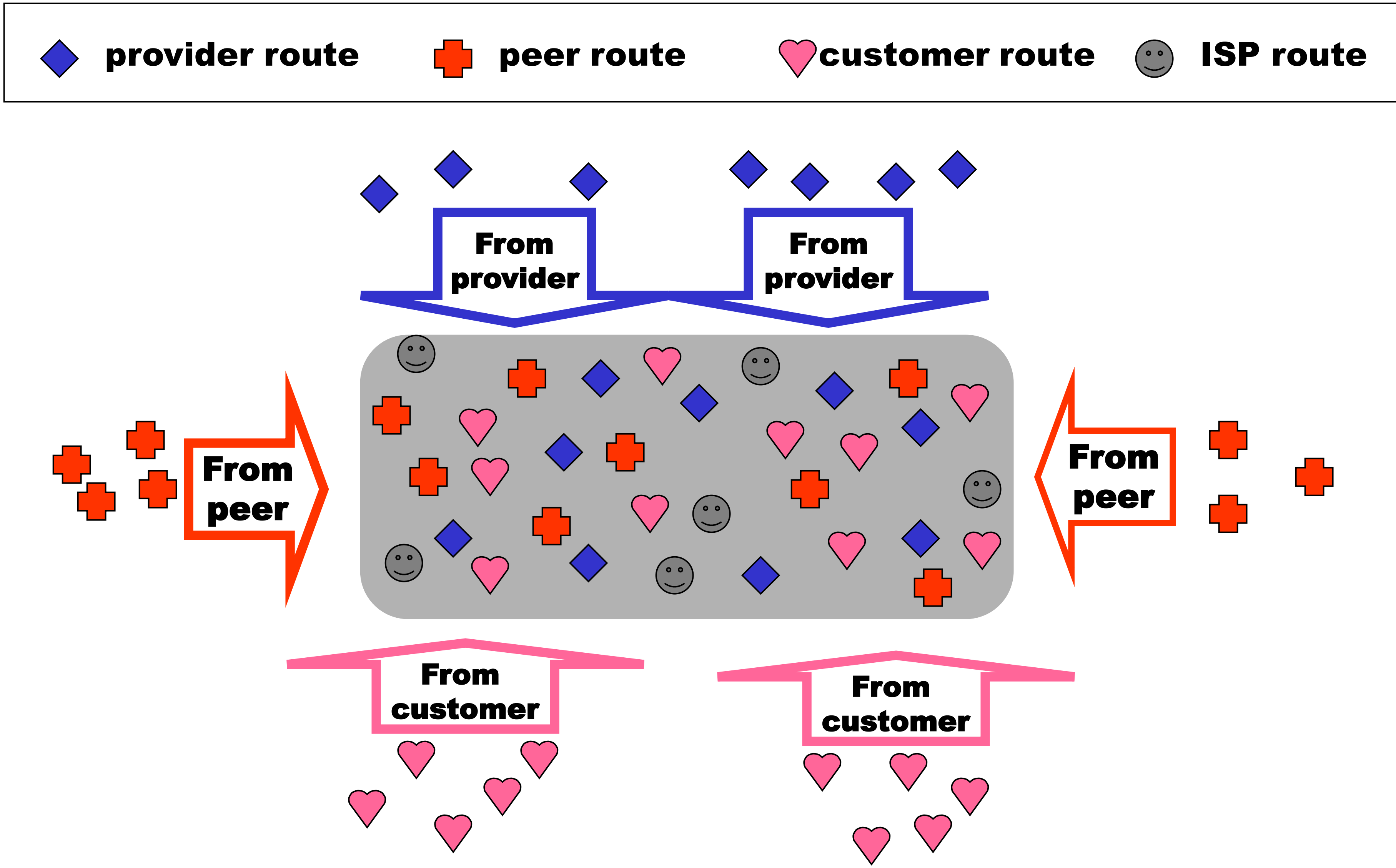
destination	next hop
135.207.0.0/16	10.10.10.10
192.0.2.0/30	10.10.10.10

Implementing Customer/Provider and Peer/Peer relationships

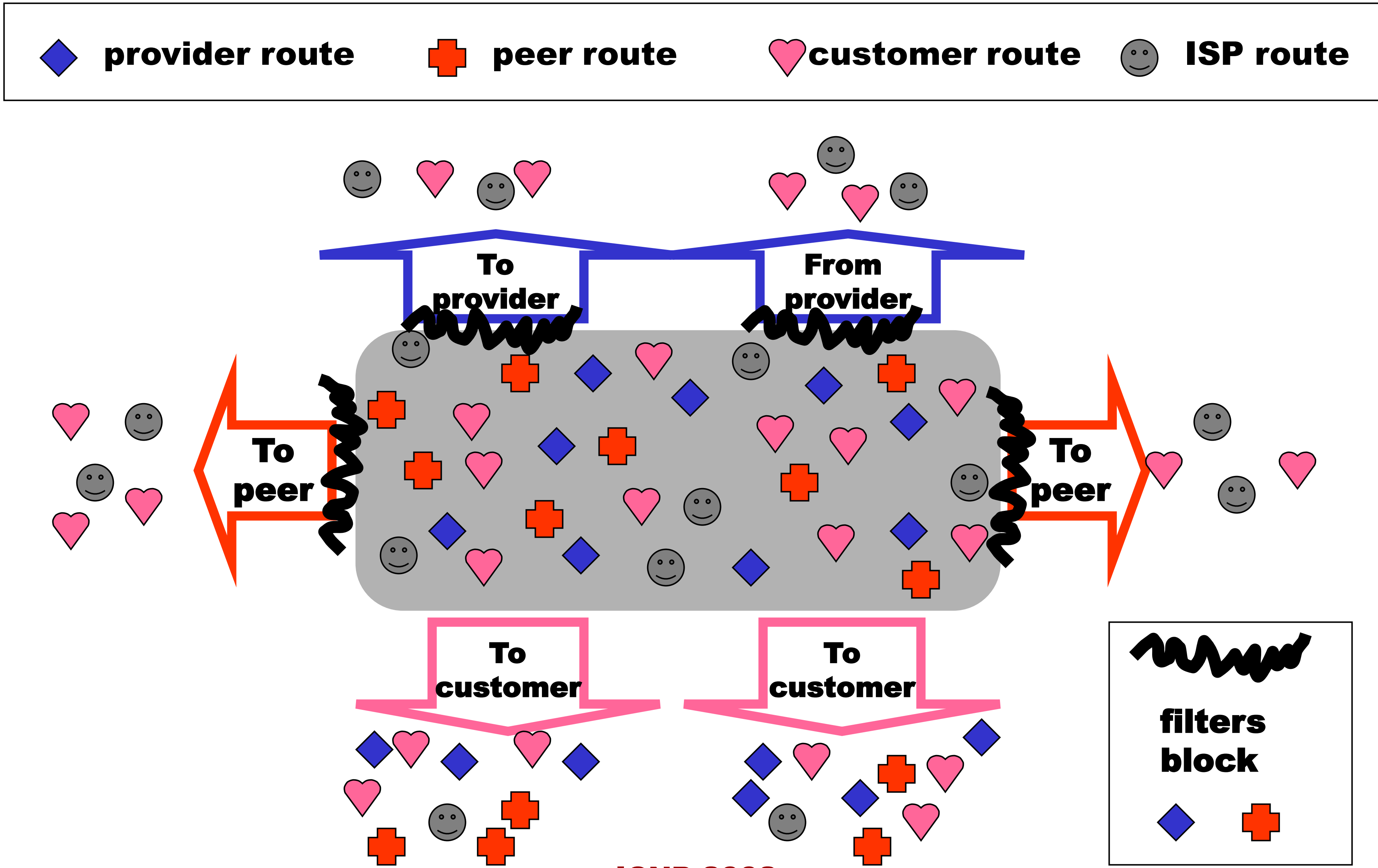
Two parts:

- **Enforce transit relationships**
  - **Outbound route filtering**
- **Enforce order of route preference**
  - **provider < peer < customer**

Import Routes

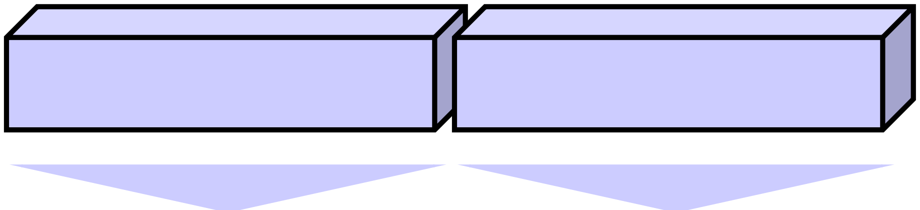


Export Routes



How Can Routes be Colored? BGP Communities!

A community value is 32 bits



By convention, first 16 bits is ASN indicating who is giving it an interpretation

community number

Used for signally within and between ASes

Very powerful BECAUSE it has no (predefined) meaning

Community Attribute = a list of community values. (So one route can belong to multiple communities)

RFC 1997 (August 1996)

Two reserved communities

no_export = 0xFFFFFF01: don’t export out of AS

no_advertise 0xFFFFFF02: don’t pass to BGP neighbors

Communities Example

- **1:100**
  - **Customer routes**
- **1:200**
  - **Peer routes**
- **1:300**
  - **Provider Routes**

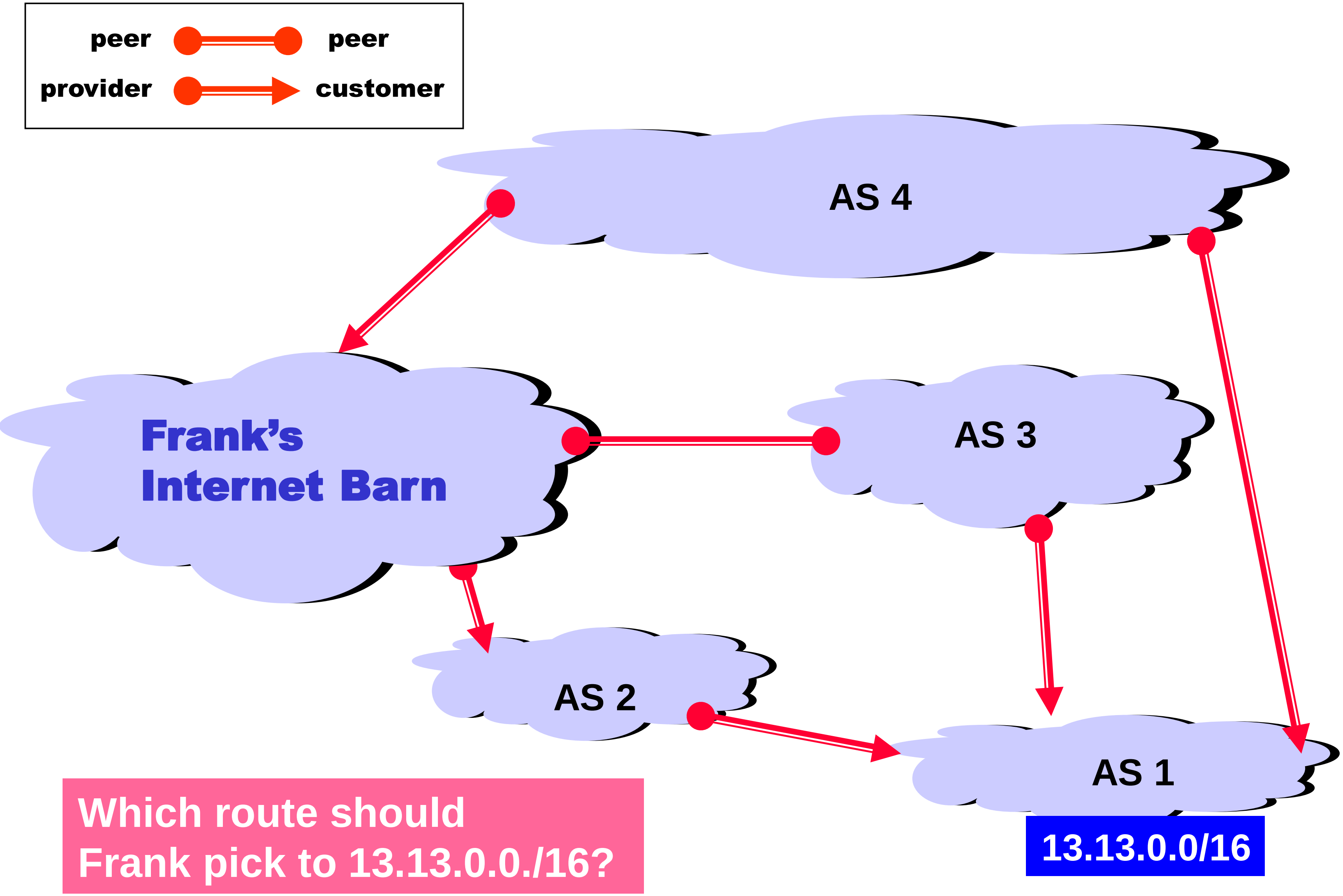
Import

- **To Customers**
  - **1:100, 1:200, 1:300**
- **To Peers**
  - **1:100**
- **To Providers**
  - **1:100**

Export

AS 1

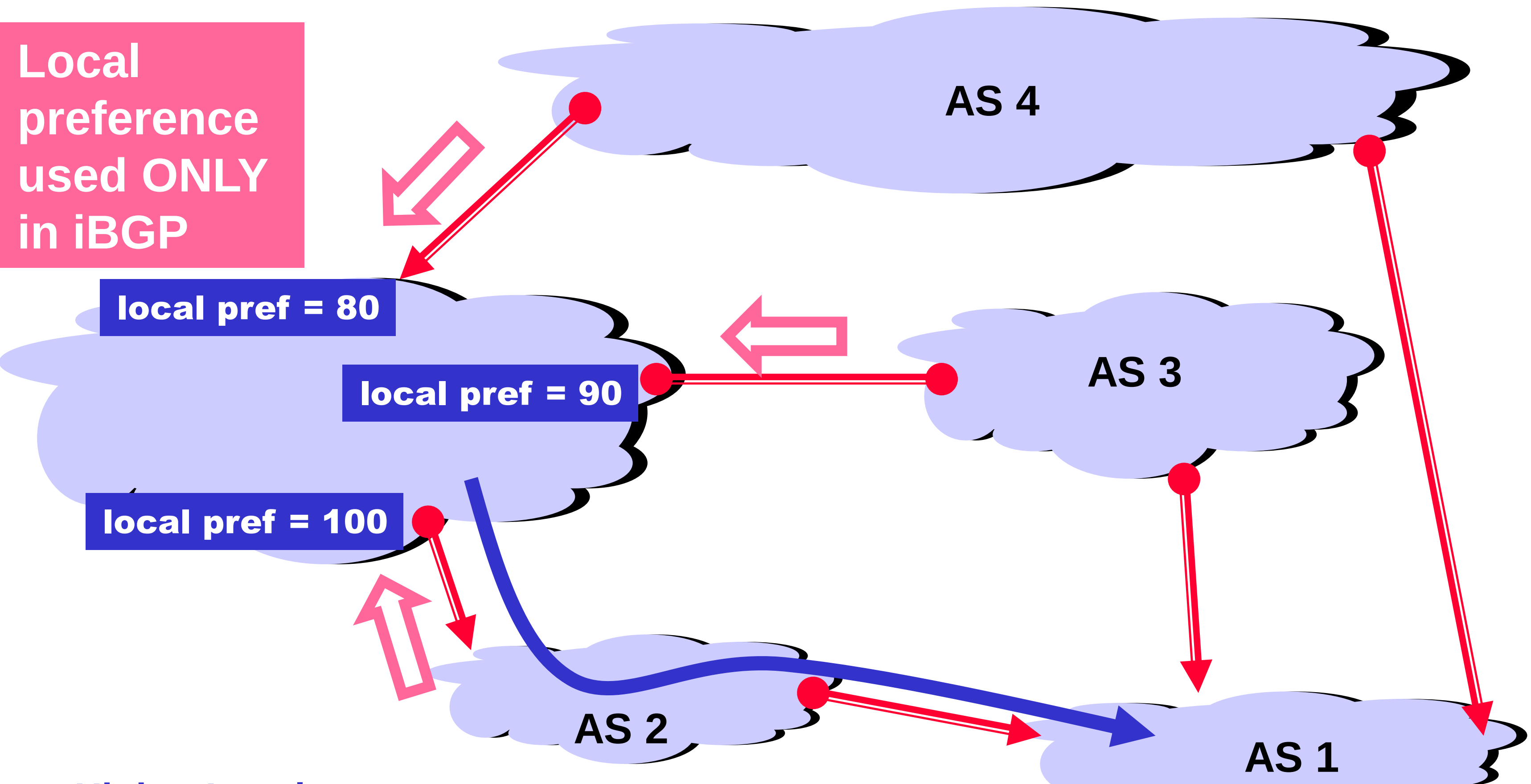
So Many Choices



Which route should
Frank pick to 13.13.0.0./16?

LOCAL PREFERENCE

Local preference used ONLY in iBGP



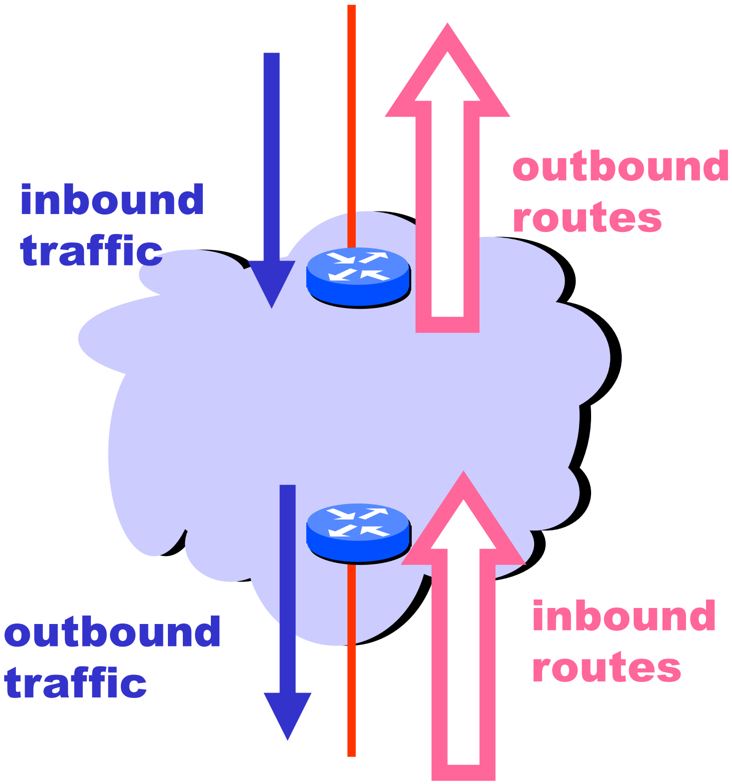
Higher Local preference values are more preferred

PART IV

Traffic Engineering with BGP

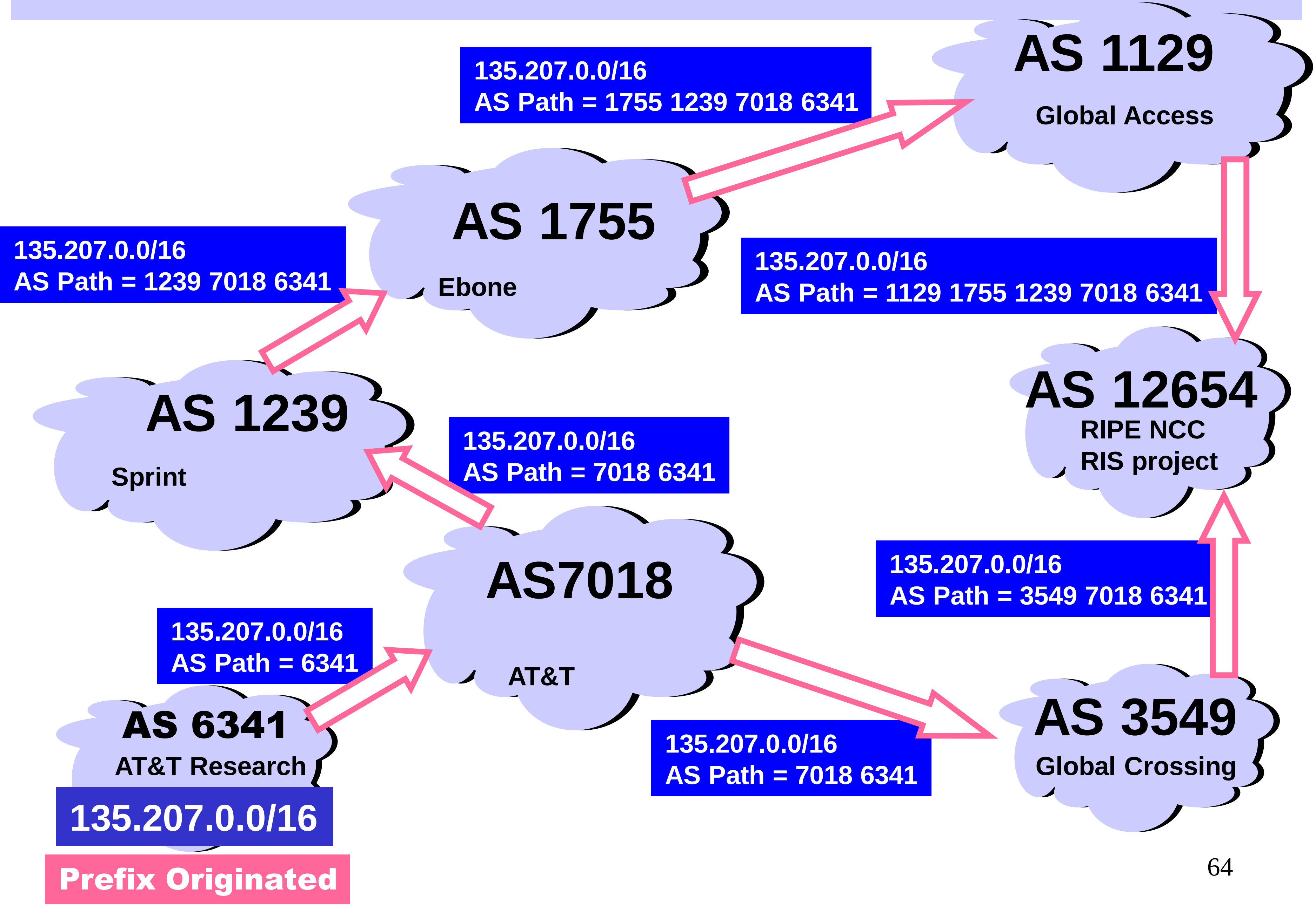
Tweak Tweak Tweak

- For <u>inbound</u> traffic
  - Filter outbound routes
  - Tweak attributes on <u>outbound</u> routes in the hope of influencing your neighbor's best route selection
- For <u>outbound</u> traffic
  - Filter <u>inbound</u> routes
  - Tweak attributes on <u>inbound</u> routes to influence best route selection



In general, an AS has more control over outbound traffic

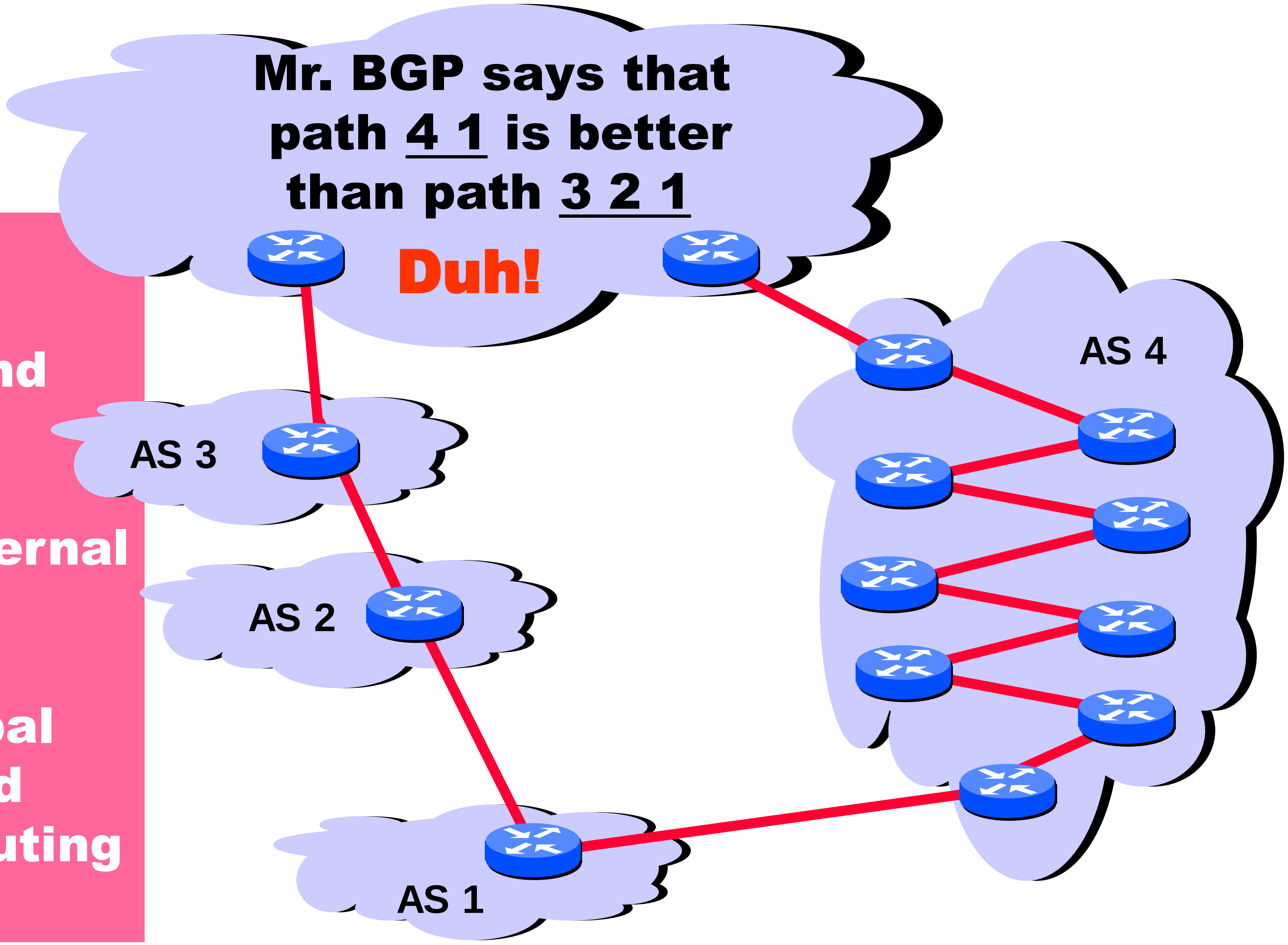
ASPATH Attribute



Shorter Doesn't Always Mean Shorter

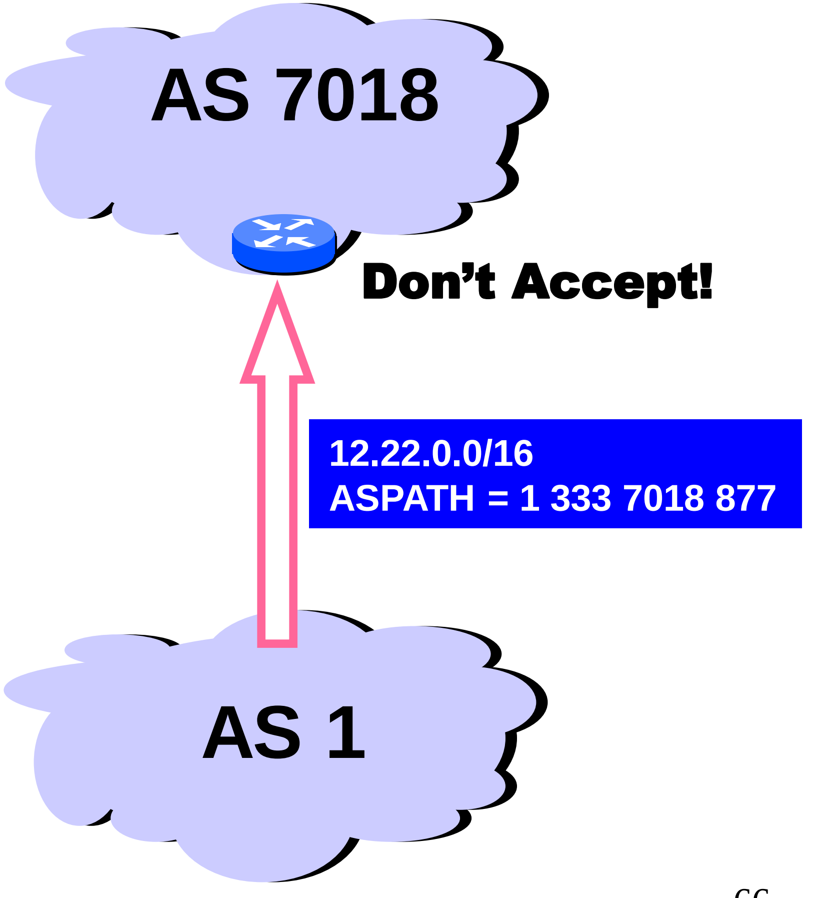
In fairness: could you do this "right" and still scale?

Exporting internal state would dramatically increase global instability and amount of routing state

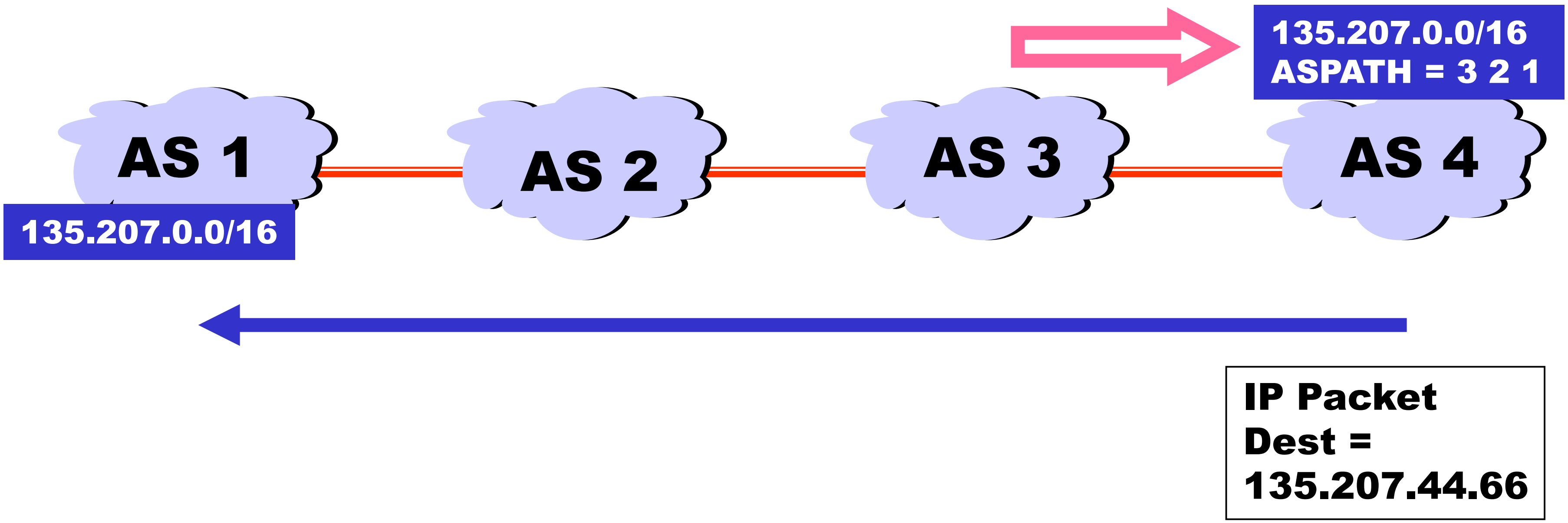


Interdomain Loop Prevention

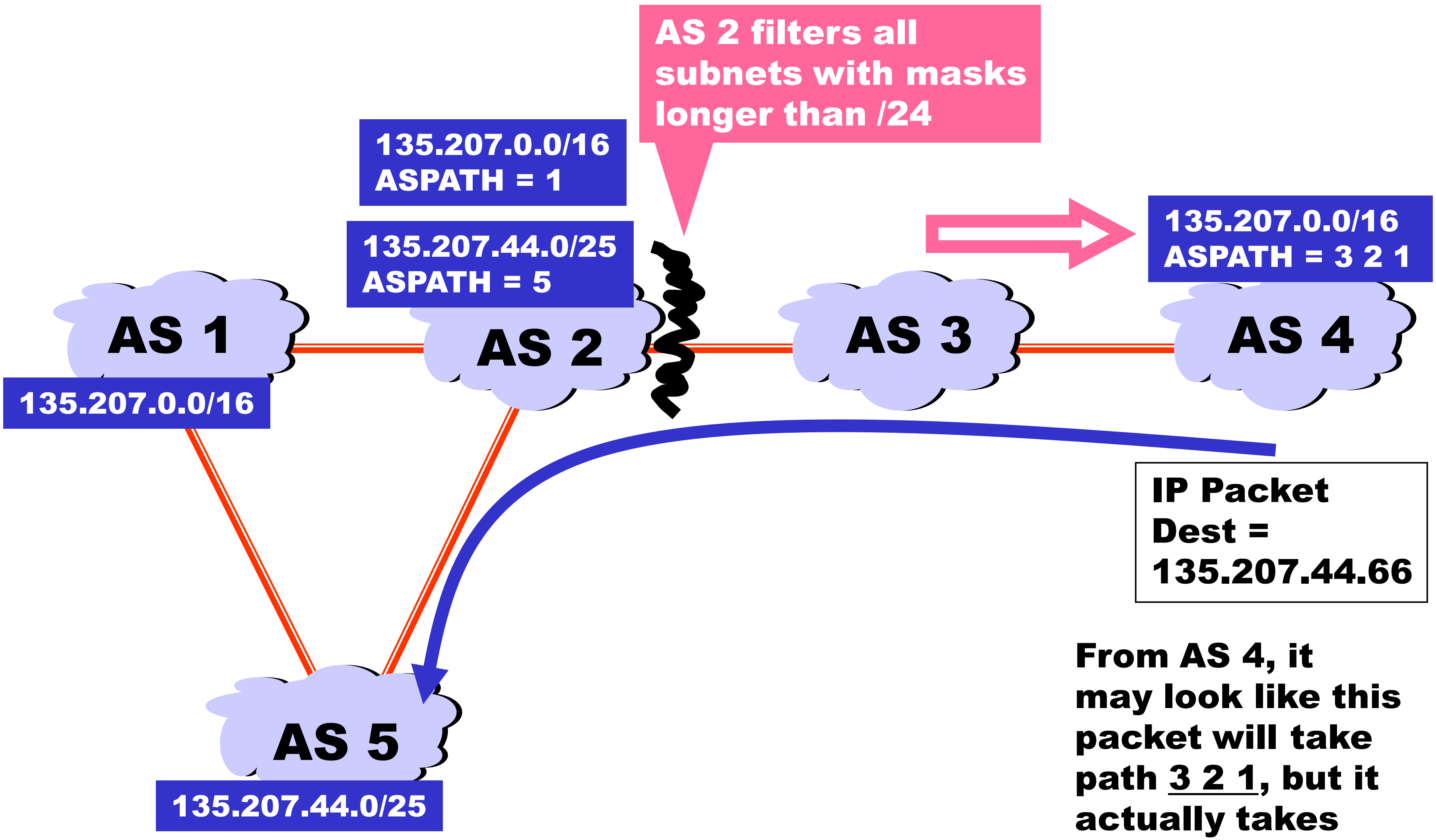
BGP at AS YYY will never accept a route with ASPATH containing YYY.



Traffic Often Follows ASPATH

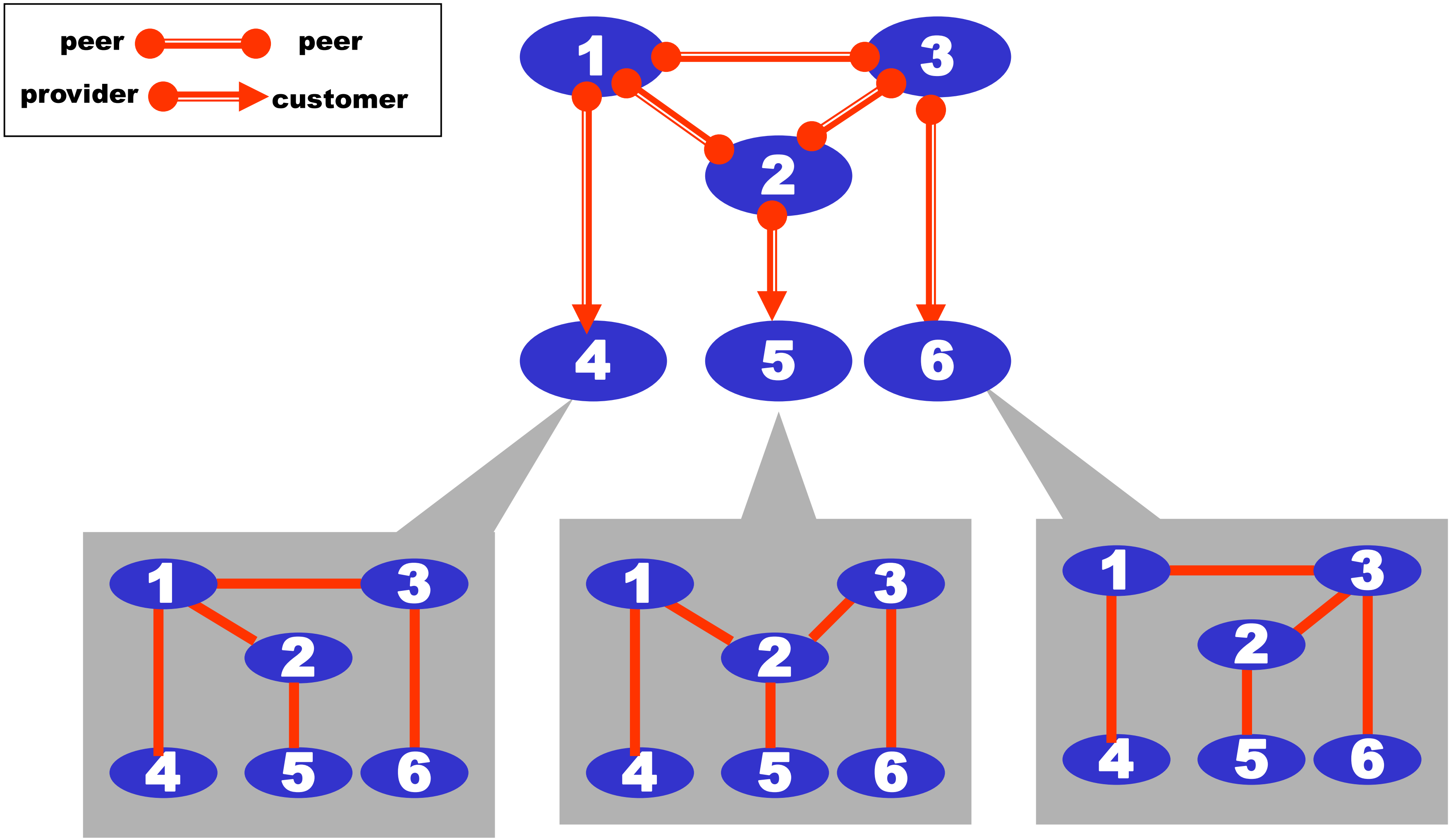


... But It Might Not

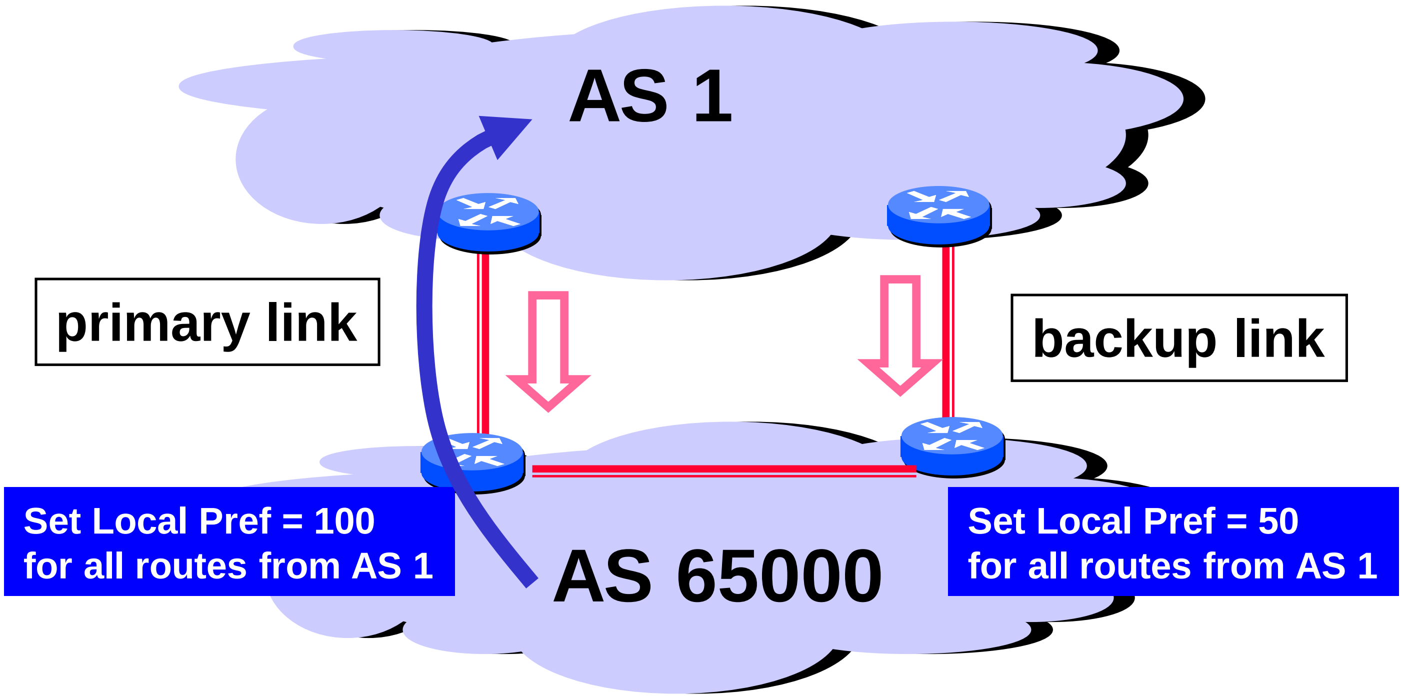


From AS 4, it may look like this packet will take path <u>3 2 1</u>, but it actually takes path <u>3 2 5</u>

AS Graphs Depend on Point of View



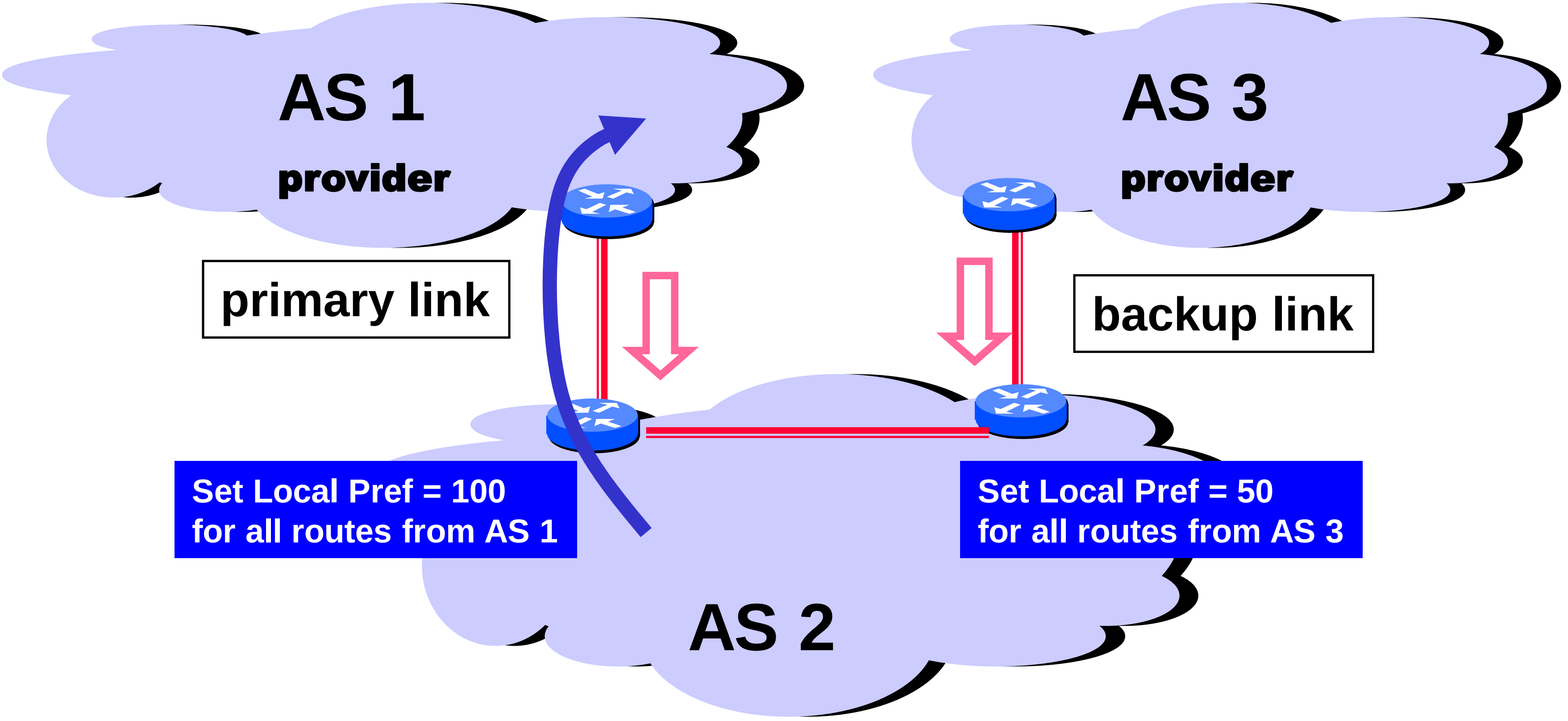
Implementing Backup Links with Local Preference (Outbound Traffic)



Forces outbound traffic to take primary link, unless link is down.

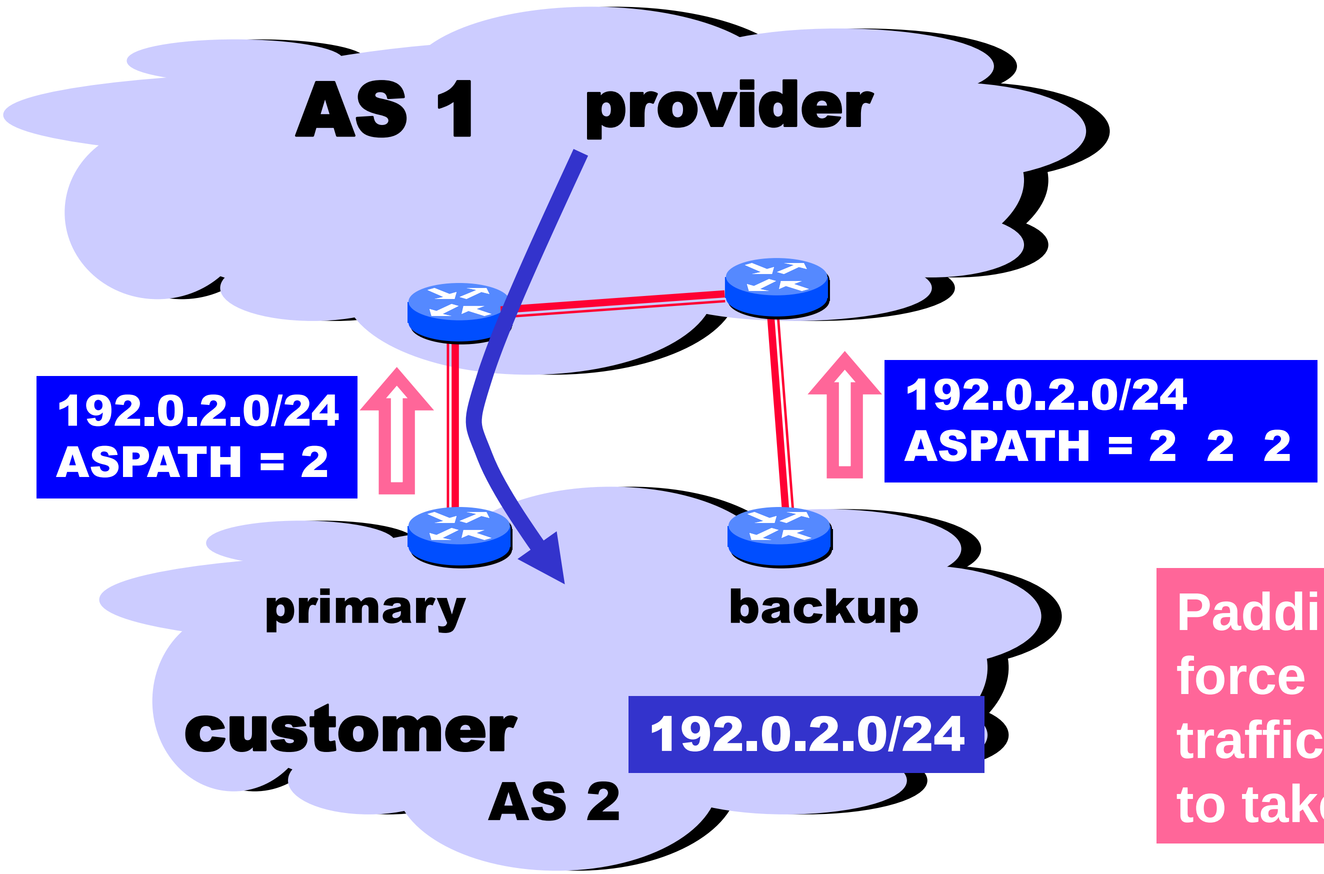
We'll talk about inbound traffic soon …

Multihomed Backups (Outbound Traffic)



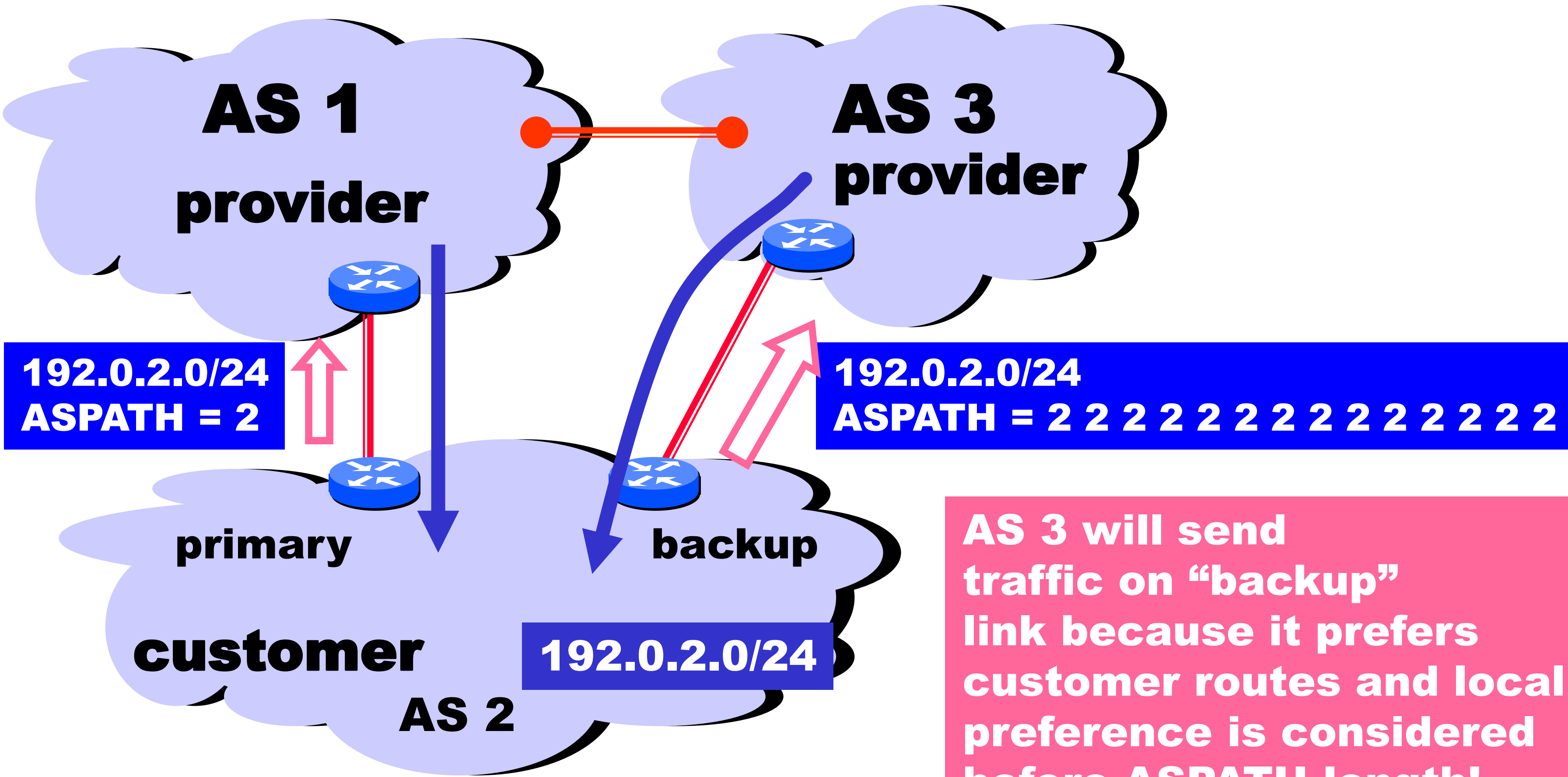
Forces outbound traffic to take primary link, unless link is down.

Shedding Inbound Traffic with ASPATH Padding. Yes, this is a Glorious Hack …



Padding will (usually) force inbound traffic from AS 1 to take primary link

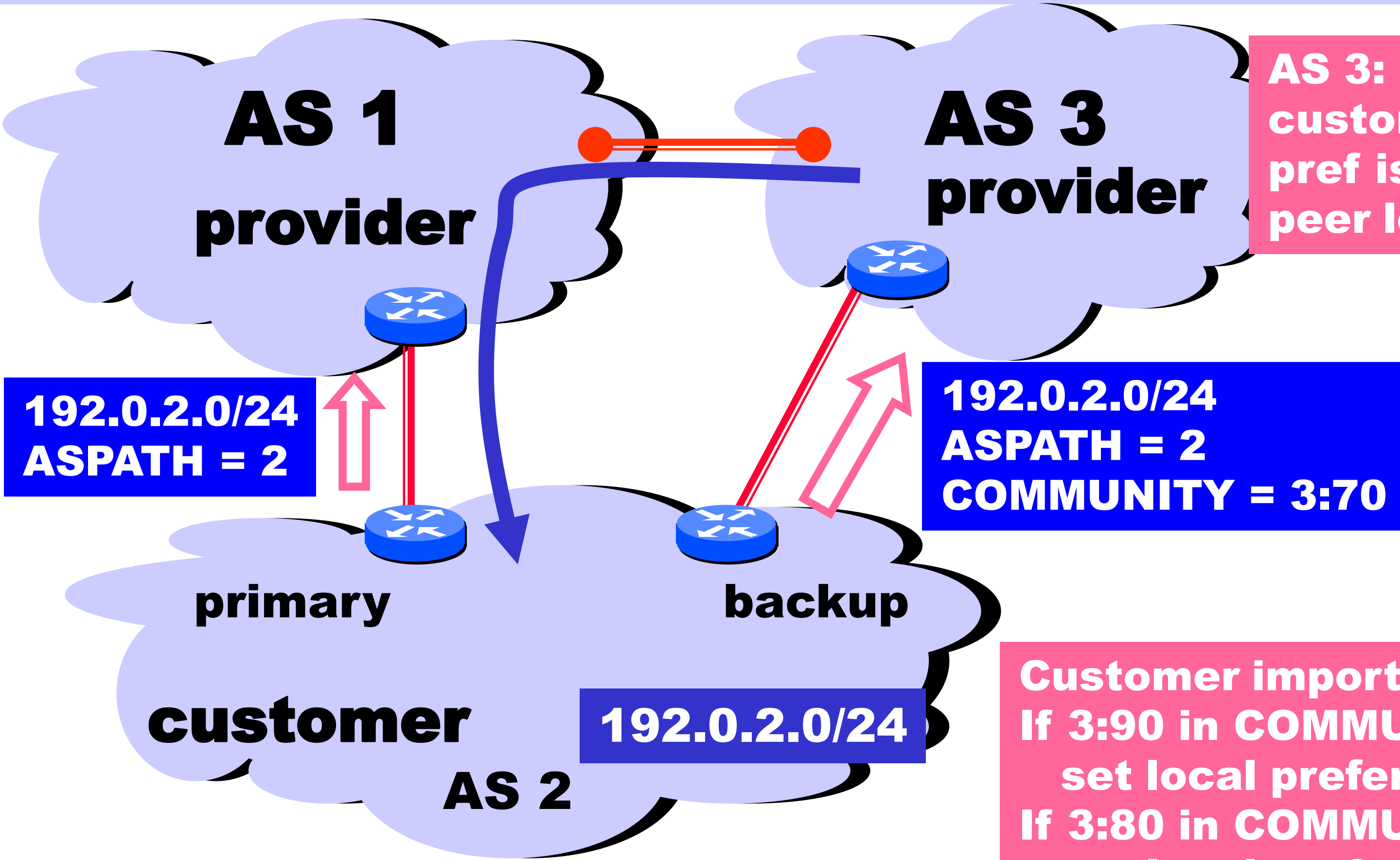
… But Padding Does Not Always Work



AS 3 will send traffic on "backup" link because it prefers customer routes and local preference is considered before ASPATH length!

Padding in this way is often used as a form of load balancing

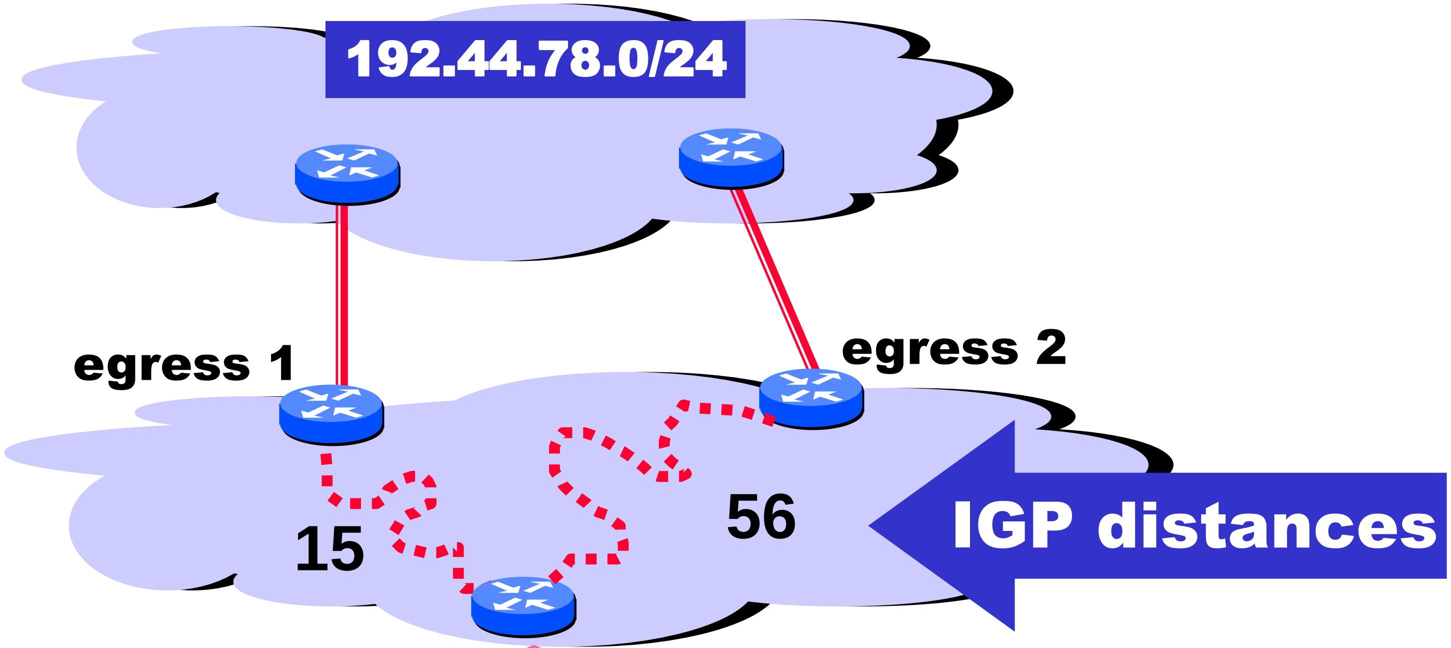
COMMUNITY Attribute to the Rescue!



AS 3: normal customer local pref is 100, peer local pref is 90

Customer import policy at AS 3:
If 3:90 in COMMUNITY then
  set local preference to 90
If 3:80 in COMMUNITY then
  set local preference to 80
If 3:70 in COMMUNITY then
  set local preference to 70

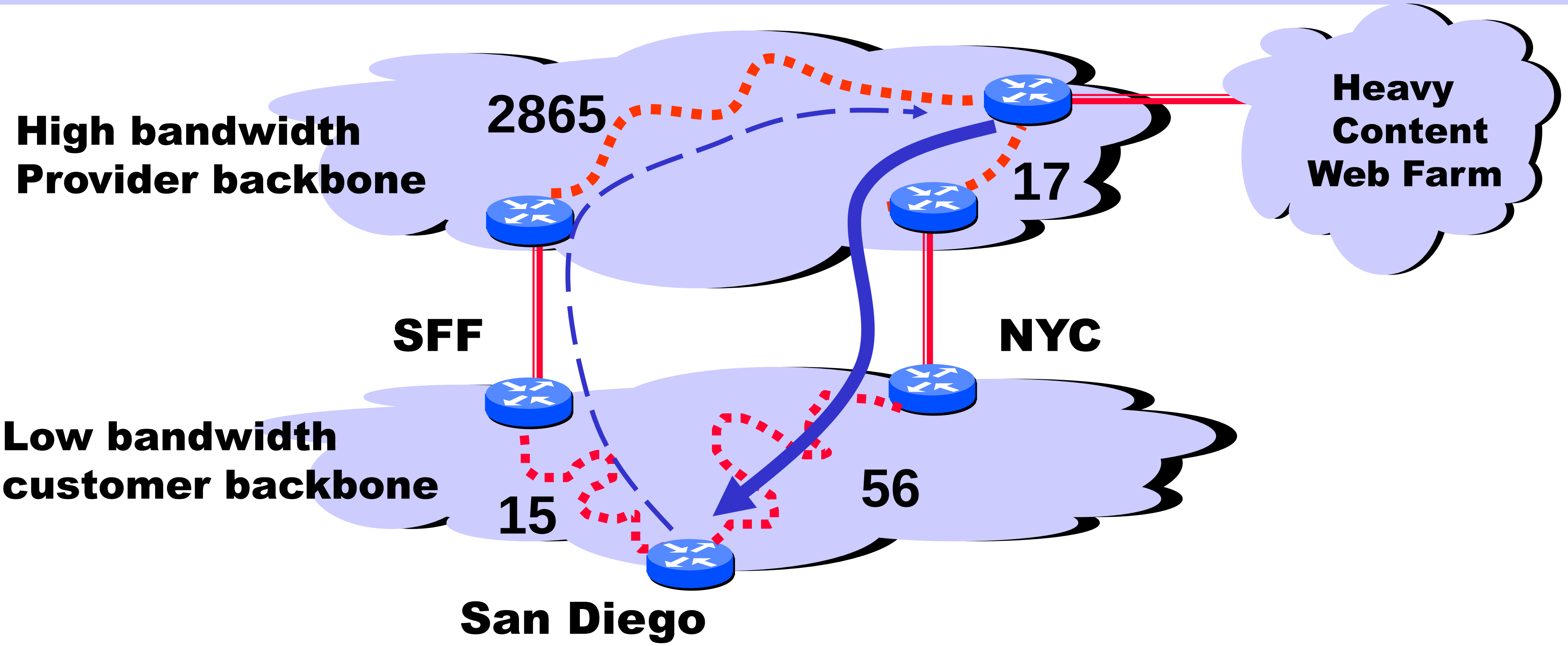
Hot Potato Routing: Go for the Closest Egress Point



This Router has two BGP routes to 192.44.78.0/24.

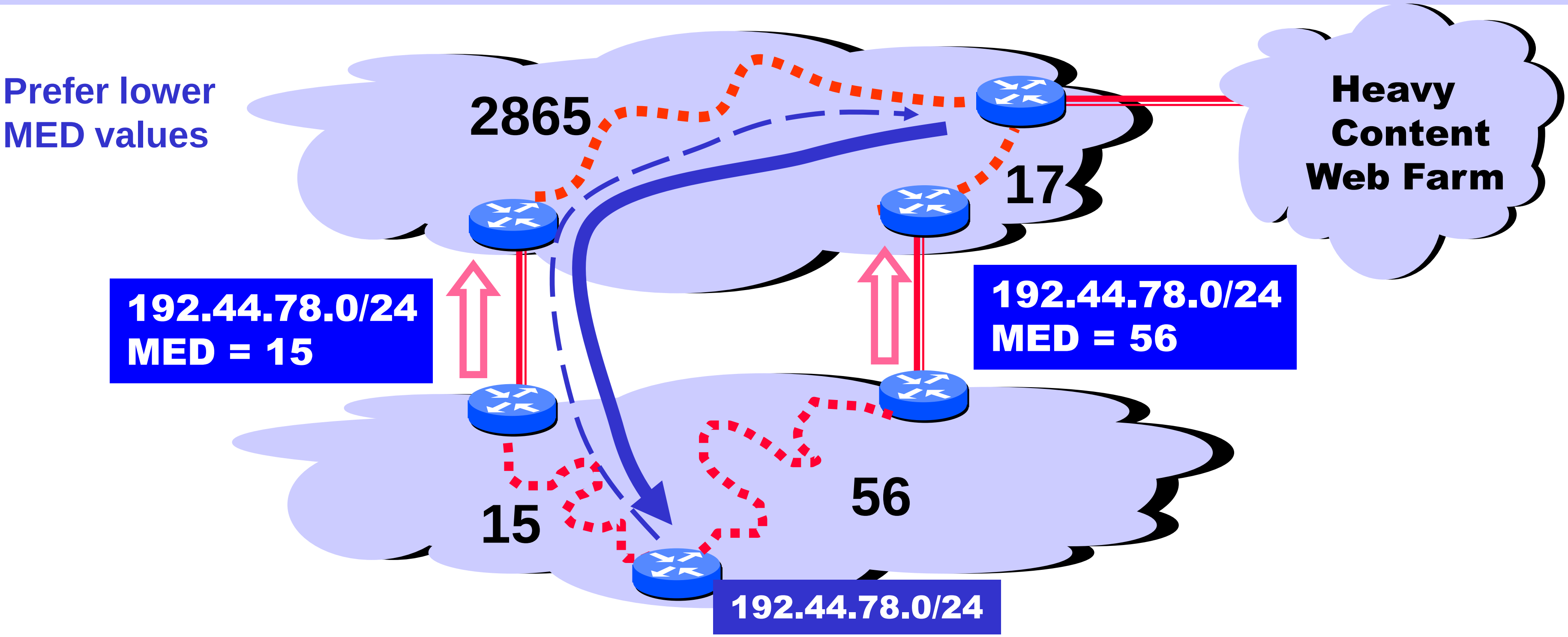
Hot potato: get traffic off of your network as Soon as possible. Go for egress 1!

Getting Burned by the Hot Potato



Many customers want their provider to carry the bits!

Cold Potato Routing with MEDs (Multi-Exit Discriminator Attribute)



This means that MEDs must be considered BEFORE IGP distance!

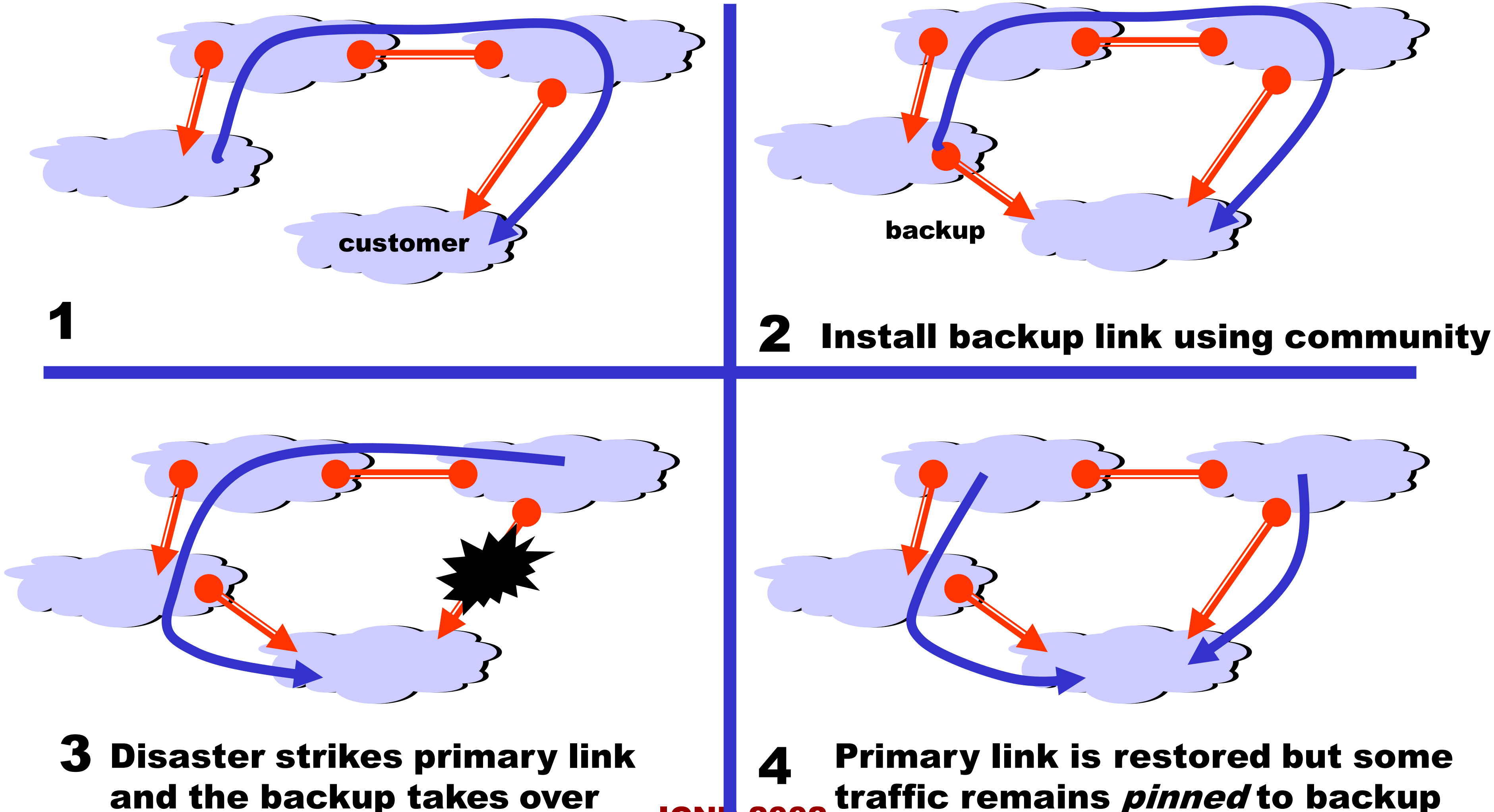
Note1 : some providers will not listen to MEDs

Note2 : MEDs need not be tied to IGP distance

PART V

A Wee Bit O' Theory: What Problem is BGP Attempting to Solve?

Policies Can Interact Strangely (“Route Pinning” Example)



News at 11:00h

- BGP is not guaranteed to converge on a stable routing. Policy interactions could lead to "livelock" protocol oscillations.

  See "Persistent Route Oscillations in Inter-domain Routing" by K. Varadhan, R. Govindan, and D. Estrin. ISI report, 1996

- Corollary: BGP is not guaranteed to recover from network failures.

What Problem is BGP Solving?

Underlying problem	Distributed means of computing a solution.
Shortest Paths	RIP, OSPF, IS-IS
X?	BGP

Separate dynamic and static semantics

static semantics

dynamic semantics

BGP Policies

BGP

Booo Hooo, Many, many complications...

Stable Paths Problem (SPP)

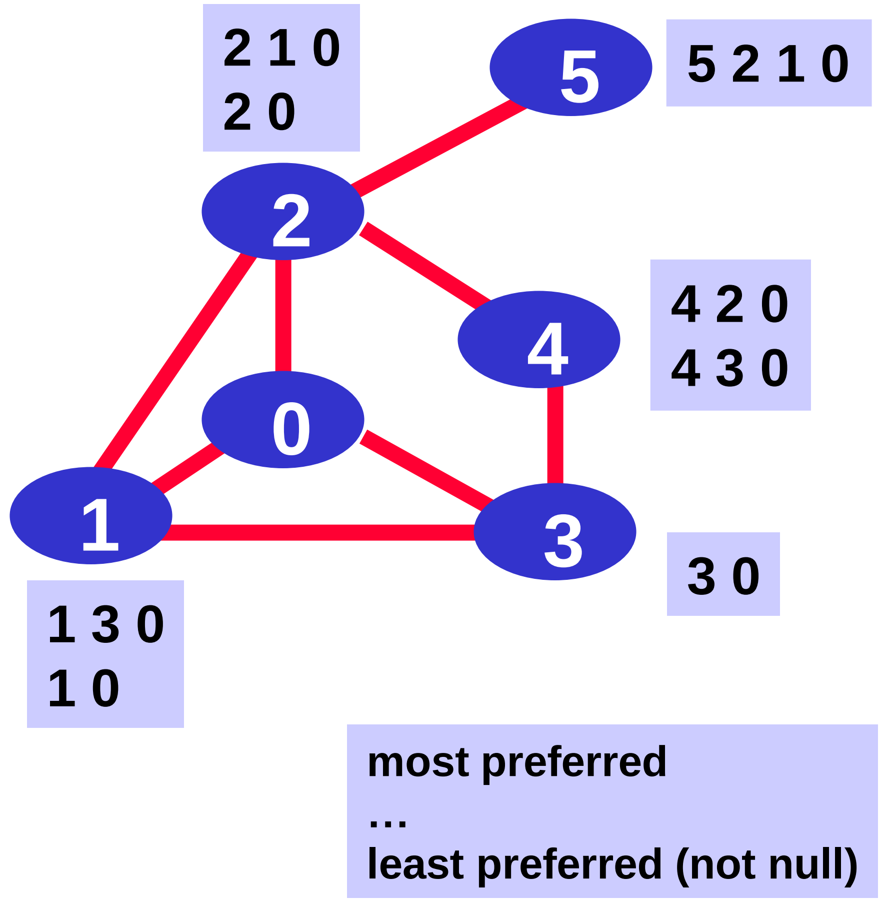
SPVP

SPVP = Simple Path Vector Protocol = a distributed algorithm for solving SPP

See [Griffin, Shepherd, Wilfong]

An instance of the *Stable Paths Problem* (SPP)

- A graph of nodes and edges,
- Node 0, called *the origin*,
- For each non-zero node, a set or permitted paths to the origin. This set always contains the "null path".
- A ranking of permitted paths at each node. Null path is always least preferred. (Not shown in diagram)



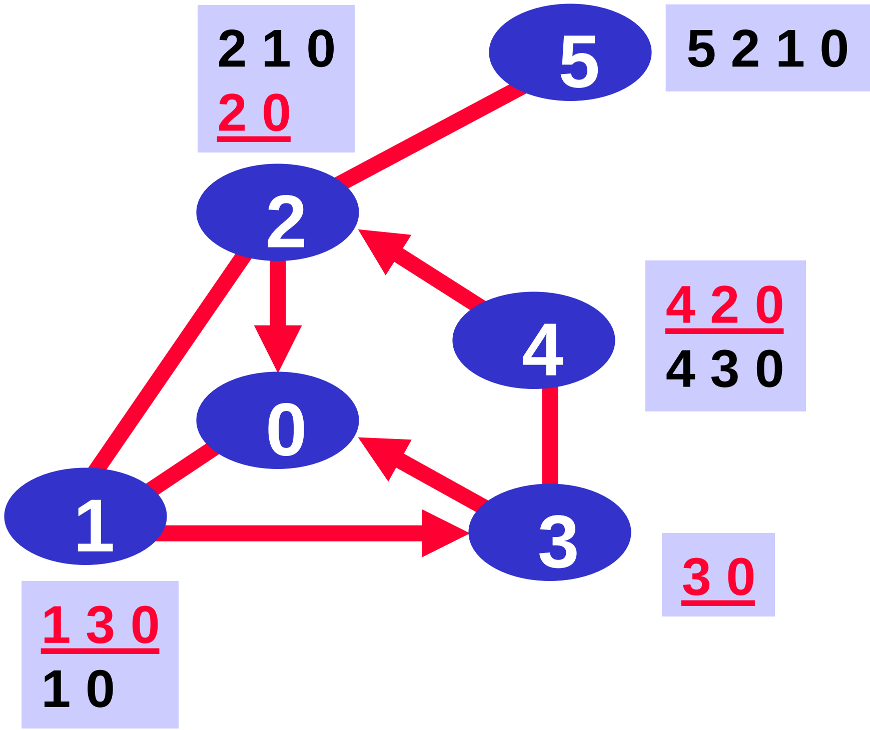
When modeling BGP : nodes represent BGP speaking routers, and 0 represents a node originating some address block

Yes, the translation gets messy!

A Solution to a Stable Paths Problem

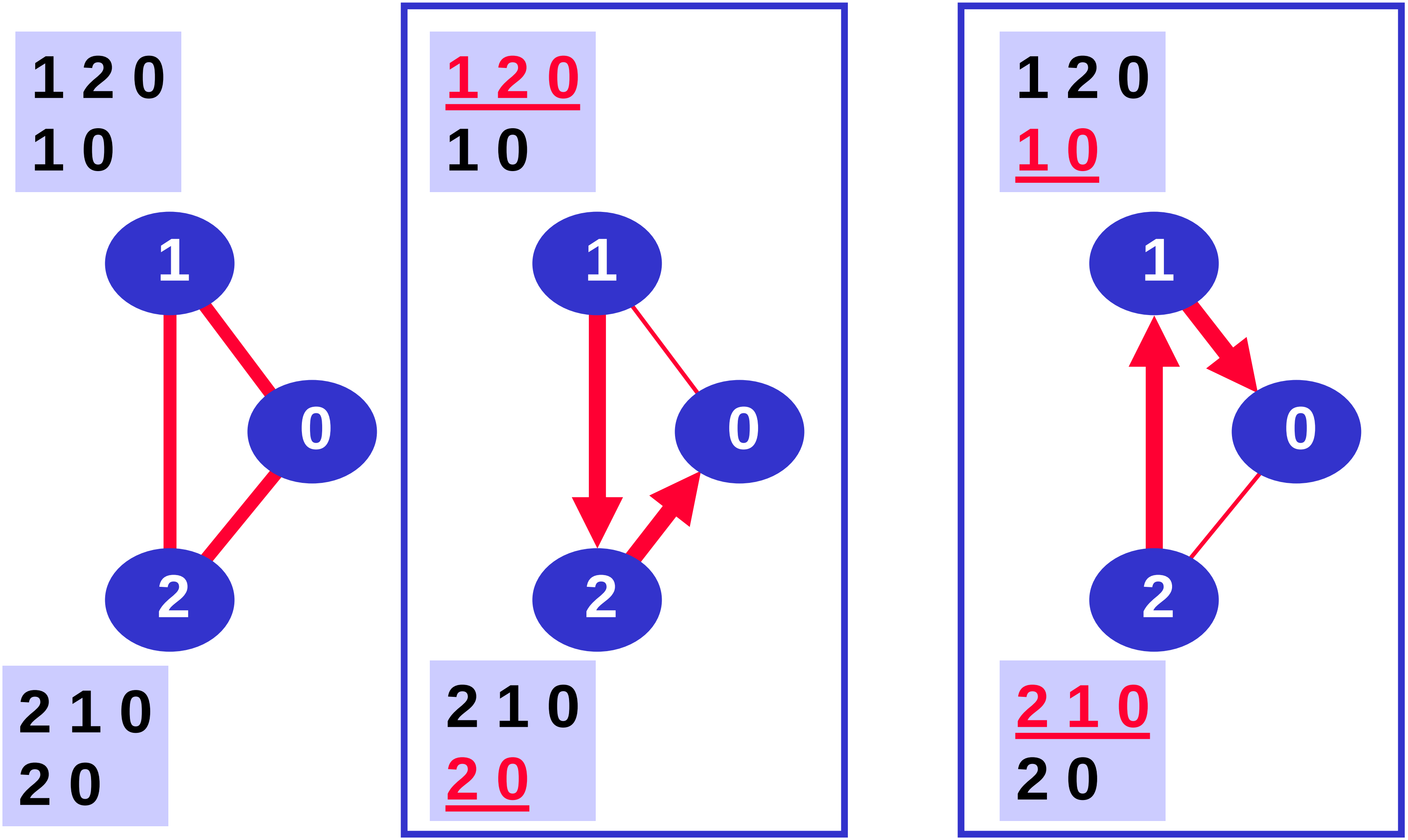
A *solution* is an assignment of permitted paths to each node such that

- node u's assigned path is either the null path or is a path uwP, where wP is assigned to node w and {u,w} is an edge in the graph,
- each node is assigned the highest ranked path among those consistent with the paths assigned to its neighbors.



A Solution need not represent a shortest path tree, or a spanning tree.

An SPP may have multiple solutions

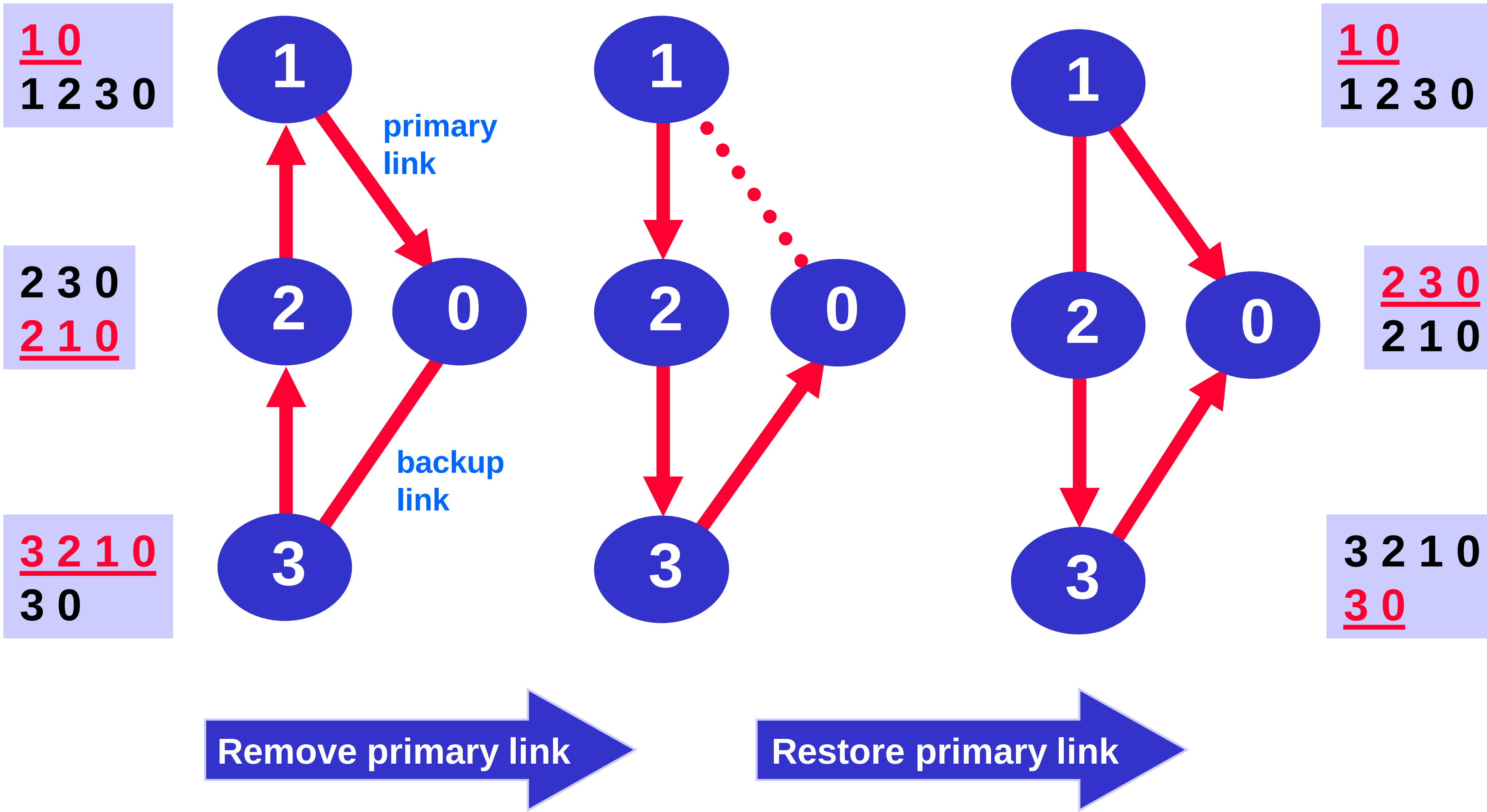


DISAGREE

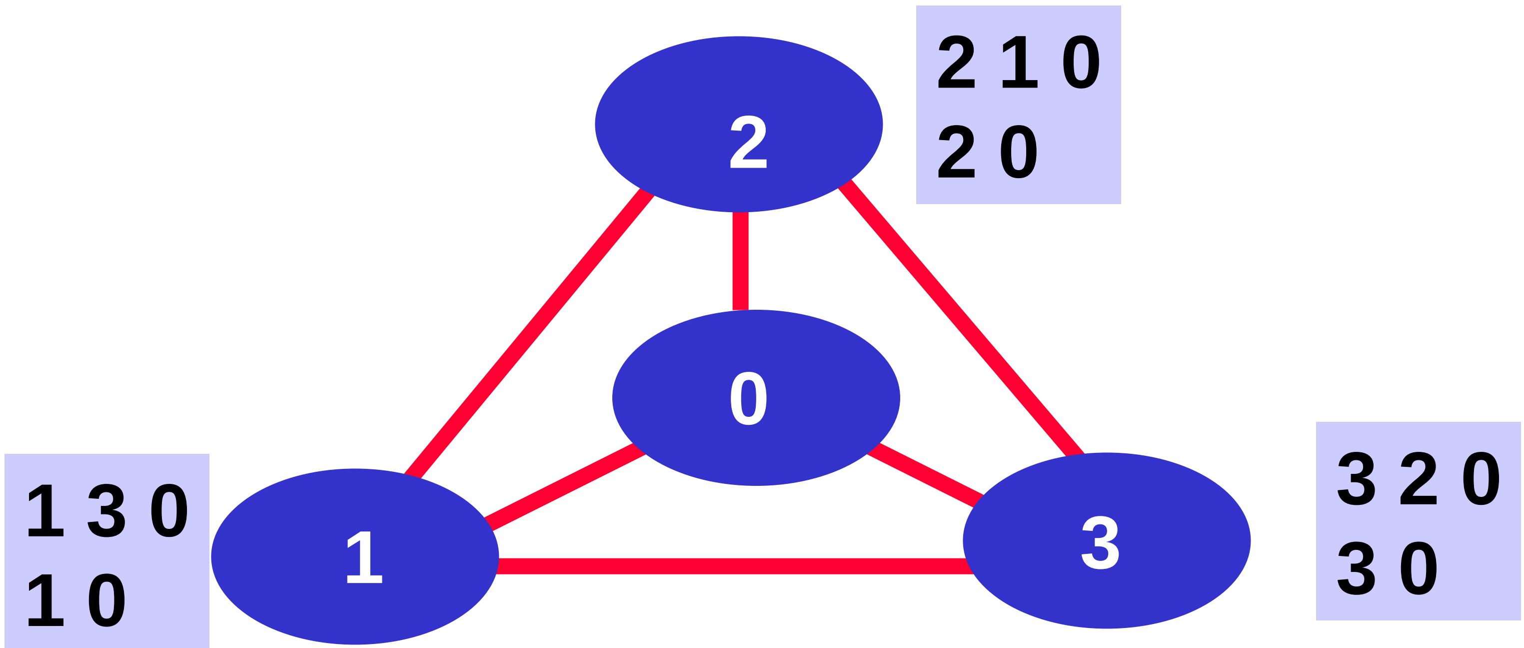
First solution

Second solution

Multiple solutions can result in "Route Triggering"



BAD GADGET : No Solution

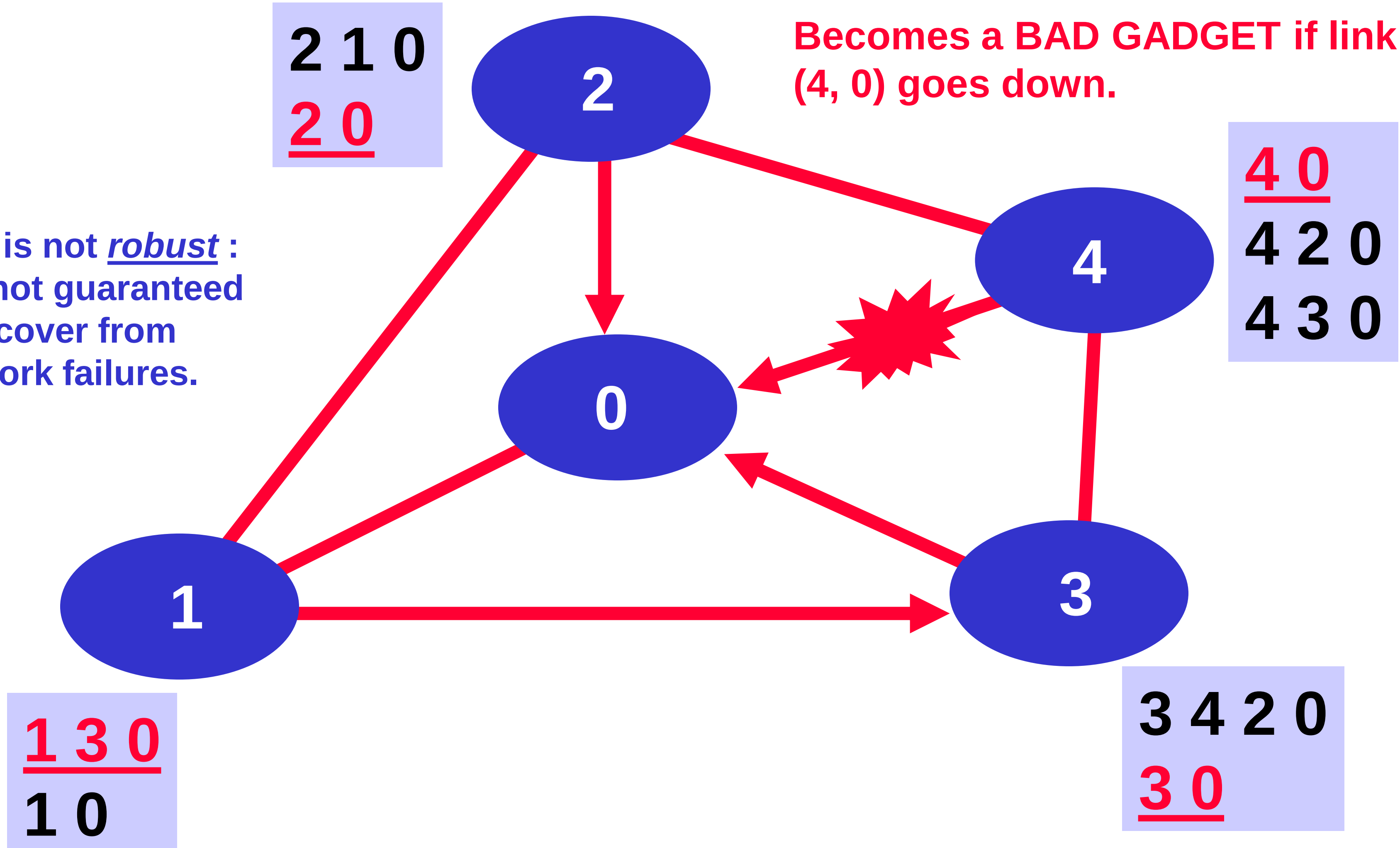


This is an SPP version of the example first presented in
Persistent Route Oscillations in Inter-Domain Routing. Kannan Varadhan, Ramesh Govindan, and Deborah Estrin. Computer Networks, Jan. 2000

SURPRISE : Beware of Backup Policies

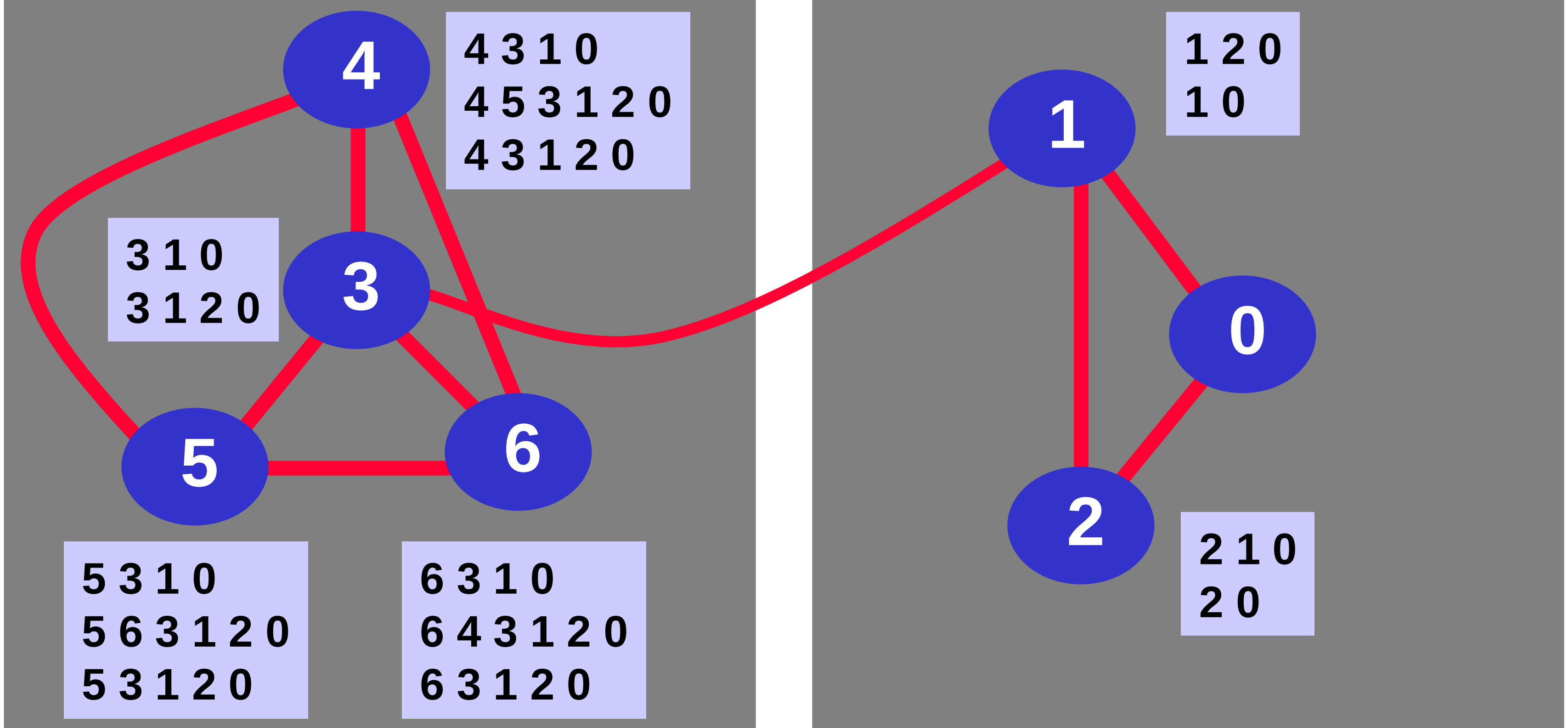
Becomes a BAD GADGET if link (4, 0) goes down.

BGP is not ***robust*** : it is not guaranteed to recover from network failures.



PRECARIOUS

Even if a solution exists, a BGP-like protocol may not find it….



This part has a solution only when node 1 is assigned the direct path (1 0).

As with DISAGREE, this part has two distinct solutions

PART VI

Current Internet Growth Trends

Large BGP Tables Considered Harmful

- Routing tables must store best routes and alternate routes
- Burden can be large for routers with many alternate routes (route reflectors for example)
- Routers have been known to die
- Increases CPU load, especially during session reset

Moore's Law may save us in theory. But in practice it means spending money to upgrade equipment …

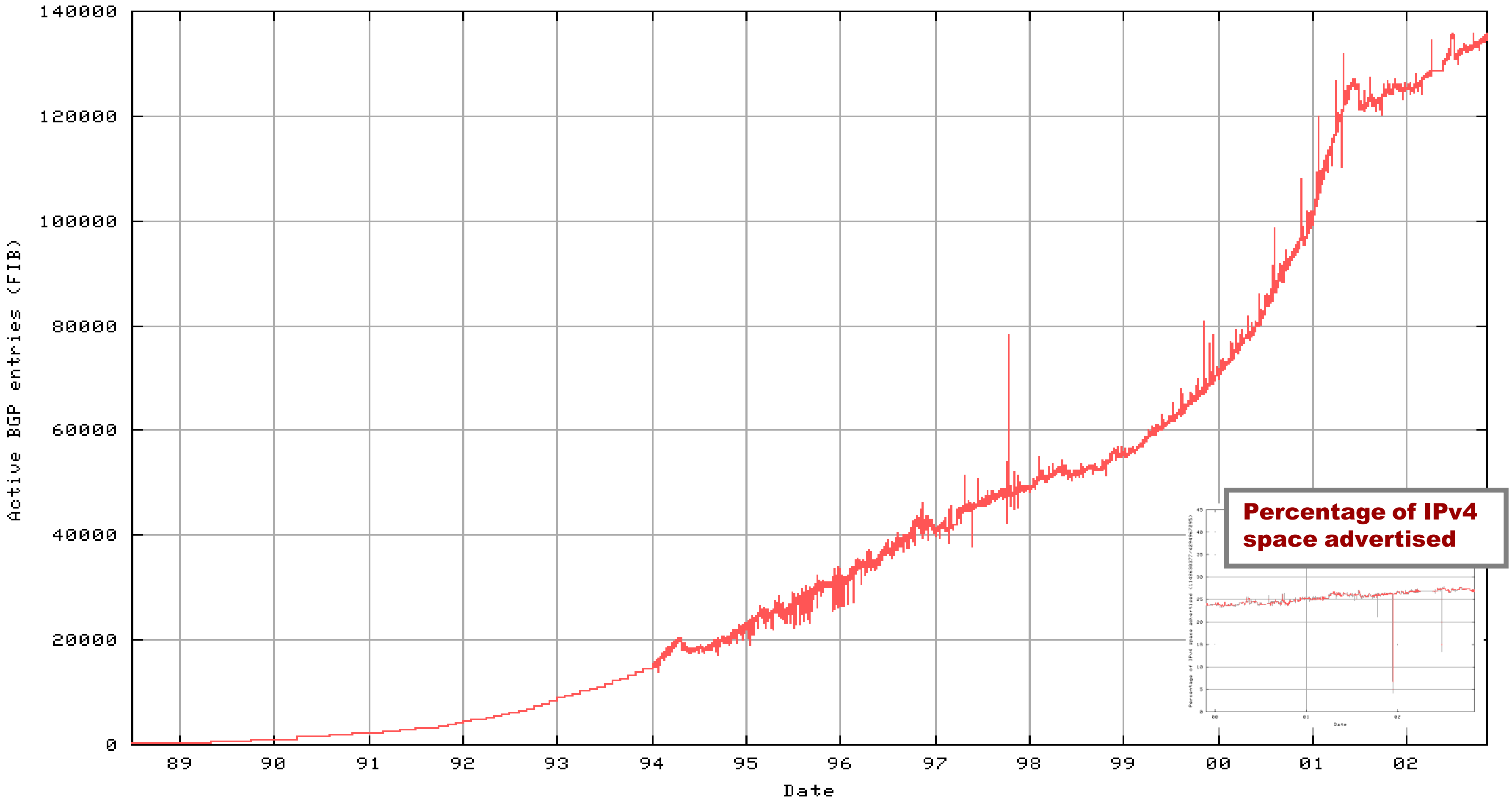
BGP Routing Tables

```
show ip bgp
BGP table version is 111849680, local router ID is 203.62.248.4
Status codes: s suppressed, d damped, h history, * valid, > best, i - internal
Origin codes: i - IGP, e - EGP, ? - incomplete

   Network          Next Hop            Metric LocPrf Weight Path

. . .
*>i192.35.25.0      134.159.0.1                    50      0 16779 1 701 703 i
*>i192.35.29.0      166.49.251.25                  50      0 5727 7018 14541 i
*>i192.35.35.0      134.159.0.1                    50      0 16779 1 701 1744 i
*>i192.35.37.0      134.159.0.1                    50      0 16779 1 3561 i
*>i192.35.39.0      134.159.0.3                    50      0 16779 1 701 80 i
*>i192.35.44.0      166.49.251.25                  50      0 5727 7018 1785 i
*>i192.35.48.0      203.62.248.34                  55      0 16779 209 7843 225 225 225 225 225 i
*>i192.35.49.0      203.62.248.34                  55      0 16779 209 7843 225 225 225 225 225 i
*>i192.35.50.0      203.62.248.34                  55      0 16779 3549 714 714 714 i
*>i192.35.51.0/25   203.62.248.34                  55      0 16779 3549 14744 14744 14744 14744 14744 14744 14744 14744 i
. . .
```

Thanks to Geoff Huston. http://www.telstra.net/ops on July 6, 2001

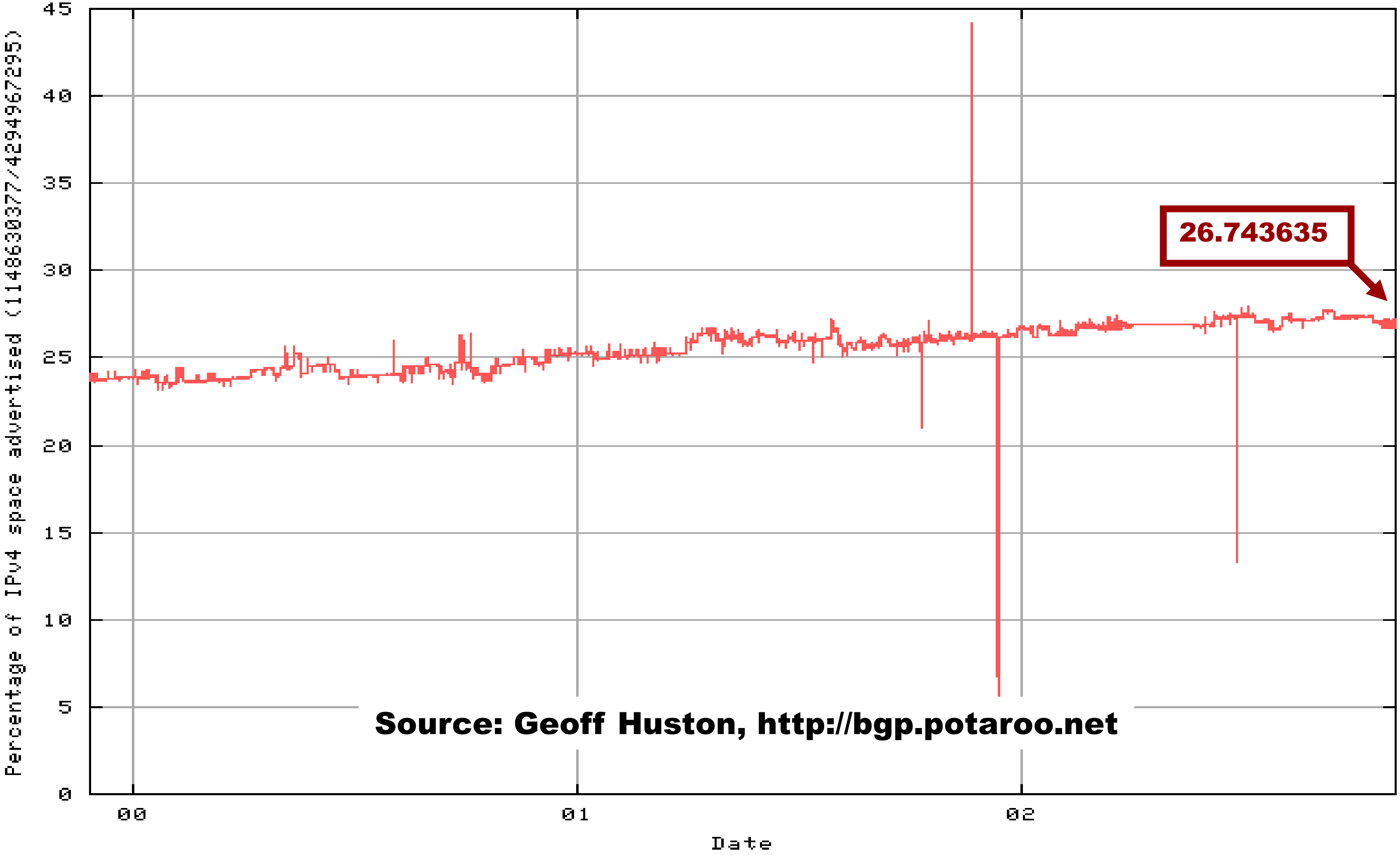
- Use “whois” queries to associate an ASN with “owner” (for example, http://www.arin.net/whois/arinwhois.html)
- 7018 = AT&T Worldnet, 701 =Uunet, 3561 = Cable & Wireless, …

Growth of BGP Routes

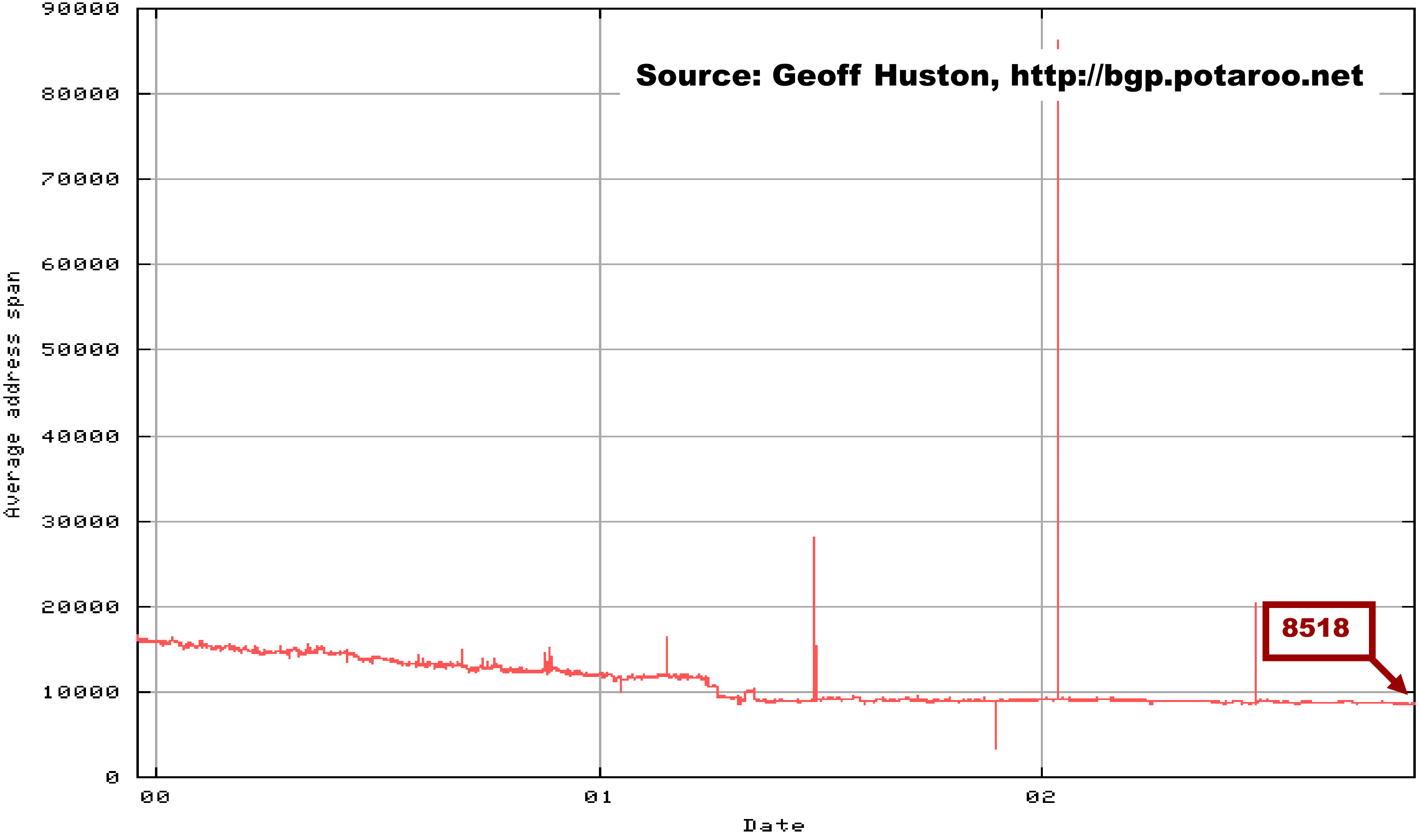


Source: Geoff Huston, http://bgp.potaroo.net, Nov. 3, 2002

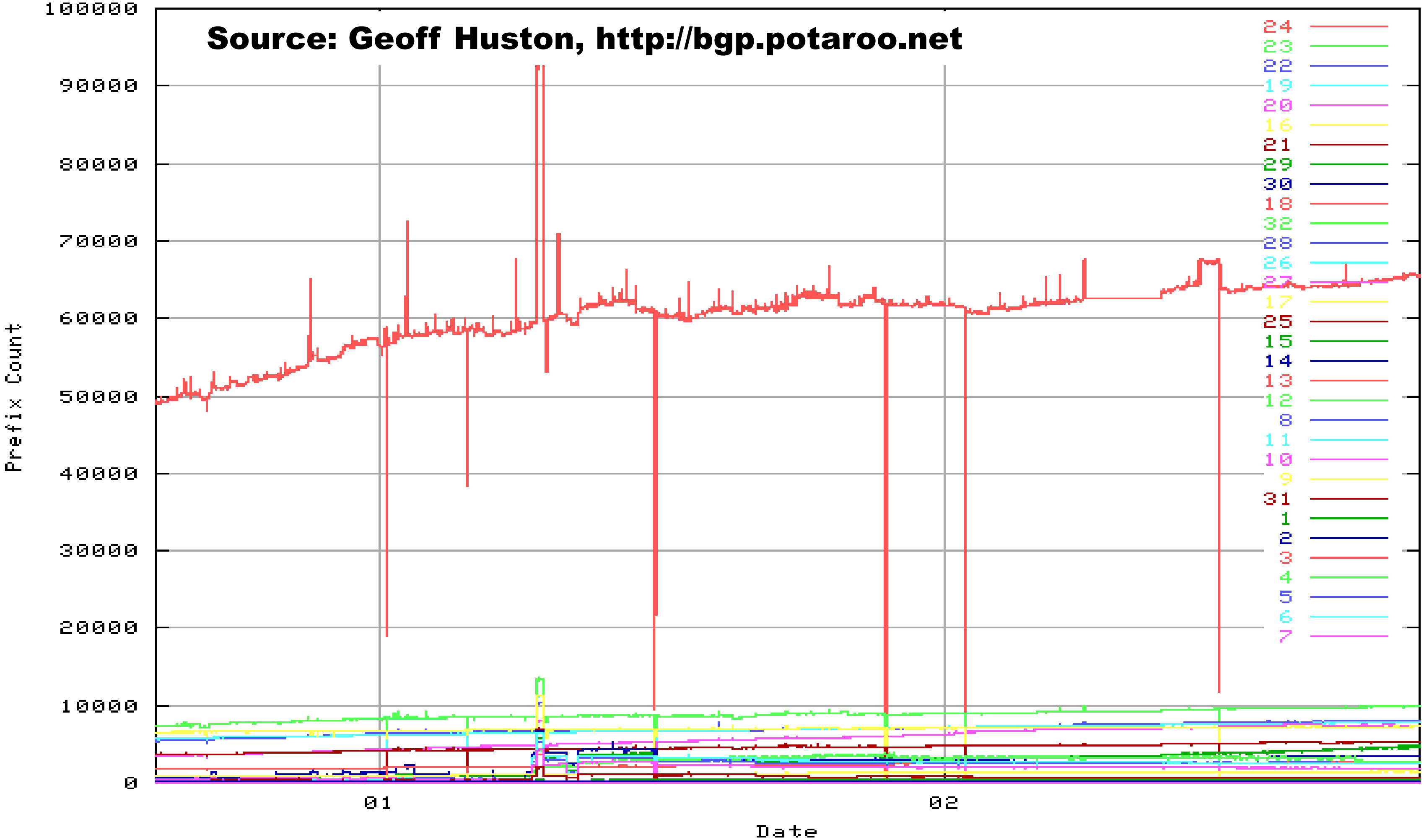
Percent of IPv4 Space Covered



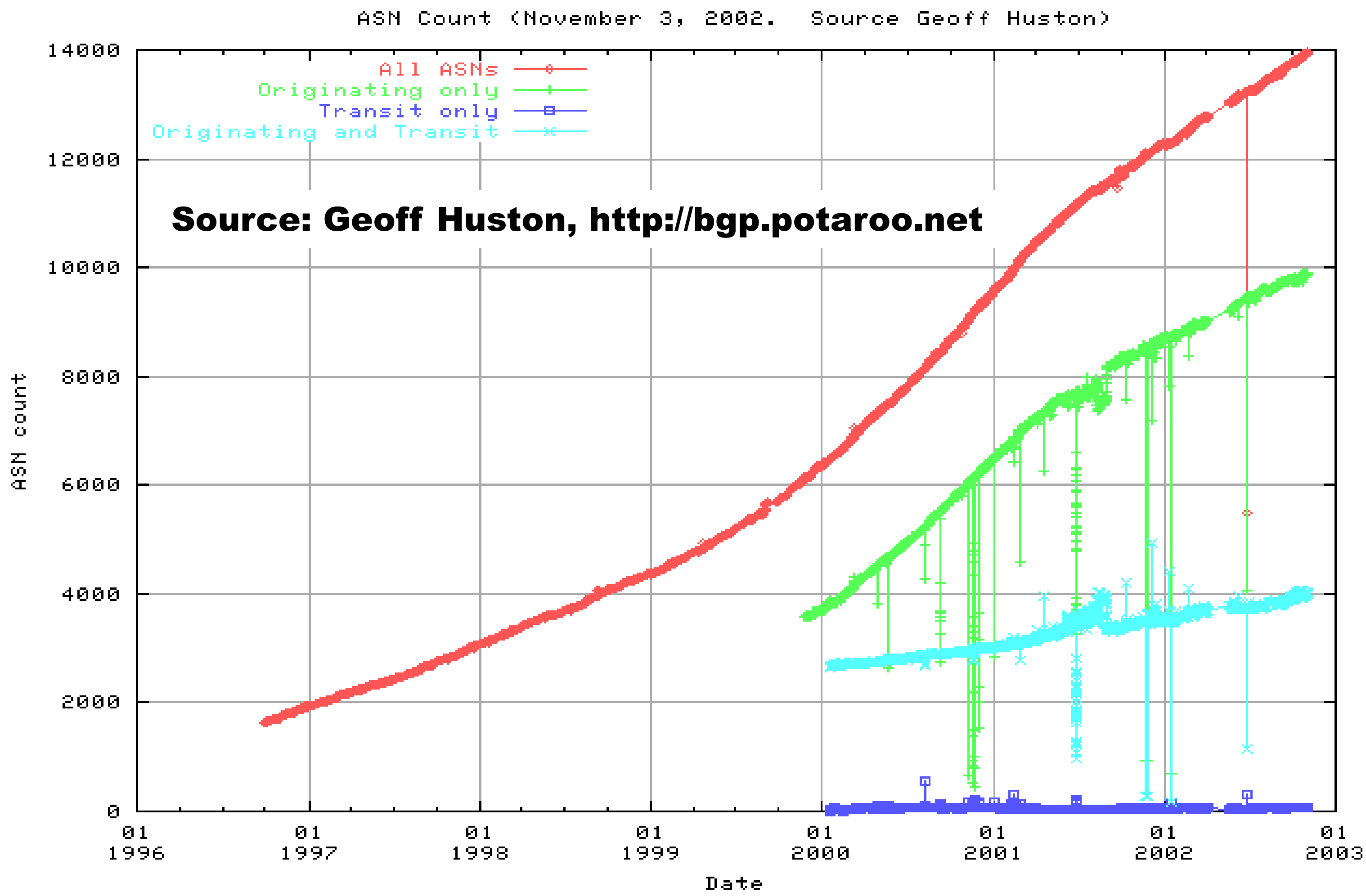
Average Span of BGP Prefixes



Prefix Lengths

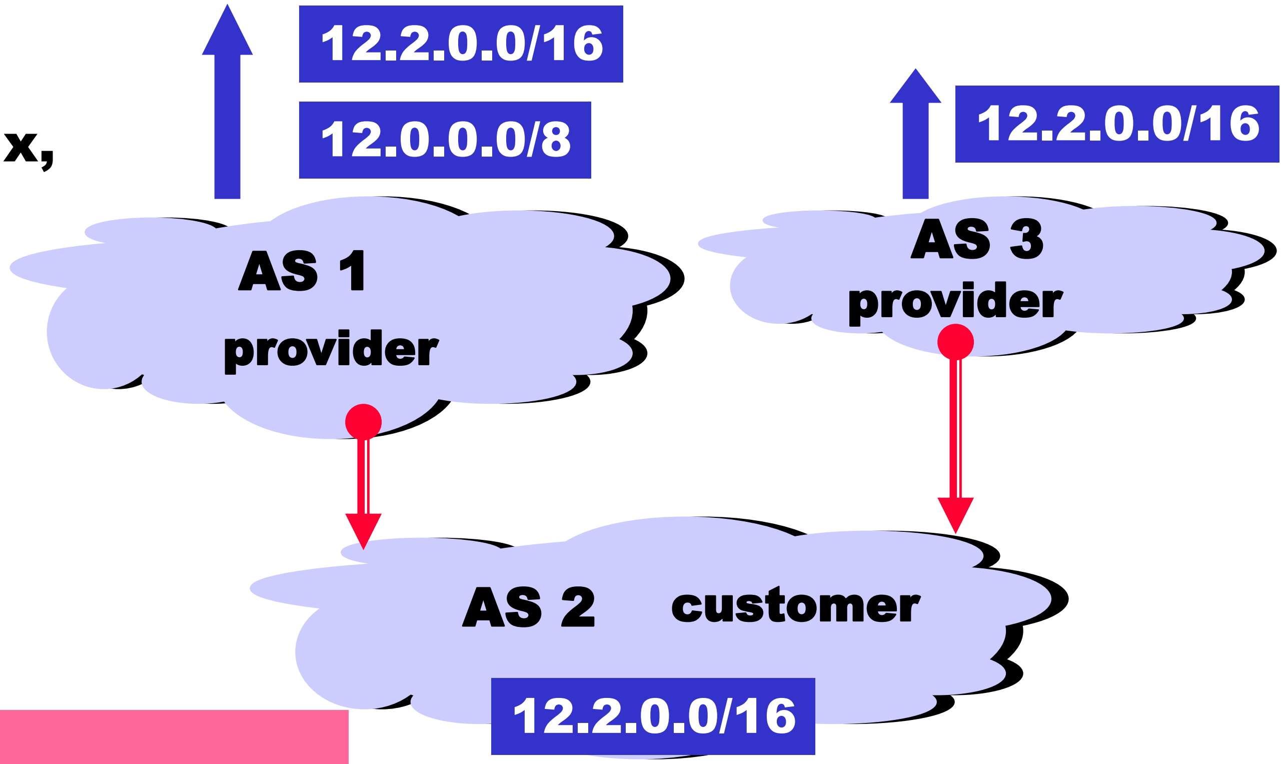


Number of Used ASNs



Deaggregation Due to Multihoming May Contribute to Table Growth

If AS 1 does not announce the more specific prefix, then most traffic to AS 2 will go through AS 3 because it is a longer match



AS 2 is "punching a hole" in The CIDR block of AS 1

For a Detailed Analysis ....

Internet Expansion, Refinement, and Churn

Andre Broido, Evi Nemeth, and kc claffy
Cooperative Association for Internet Data Analysis - CAIDA
San Diego Supercomputer Center,
University of California, San Diego

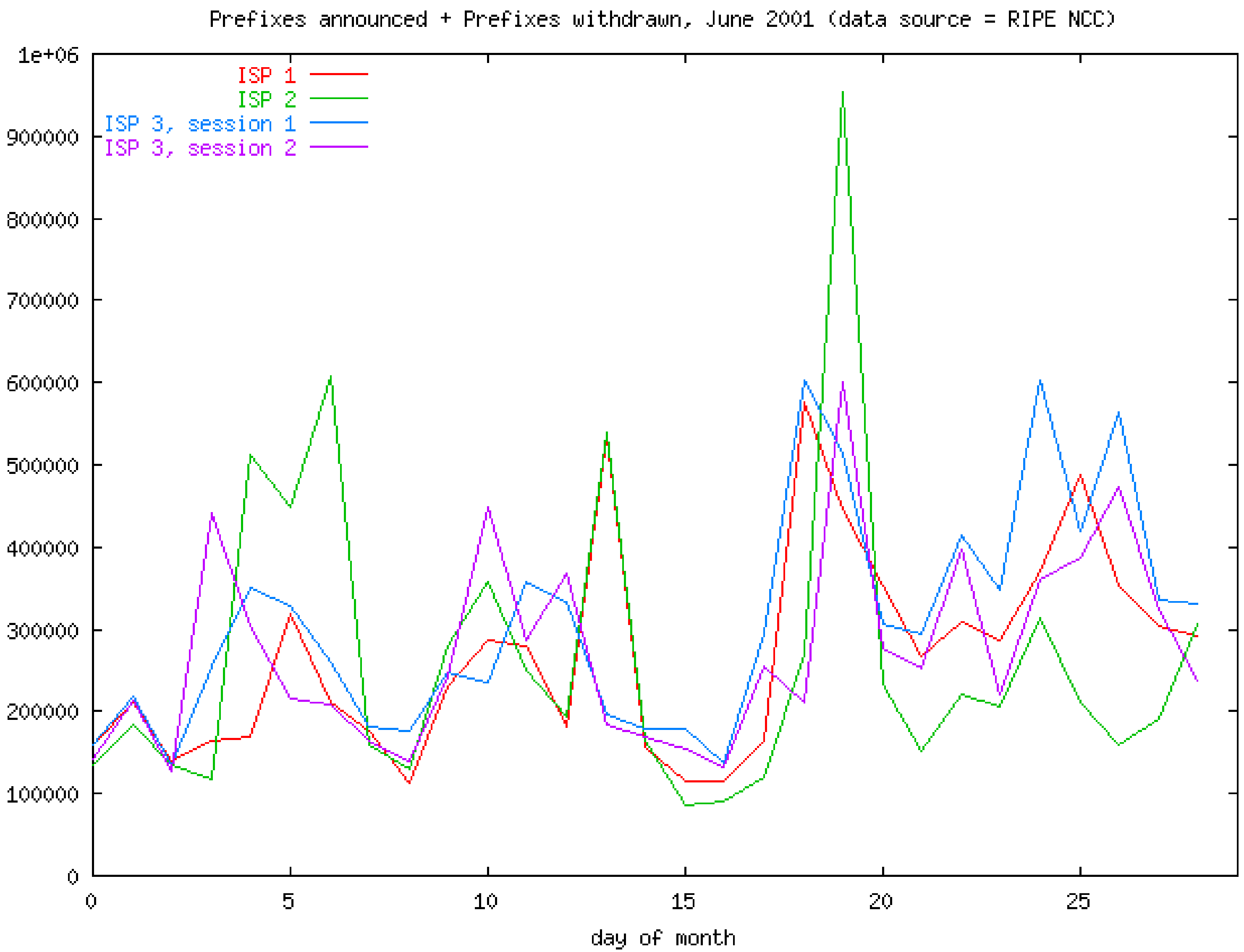
http://www.caida.org/outreach/papers/2002/EGR/

BGP Dynamics

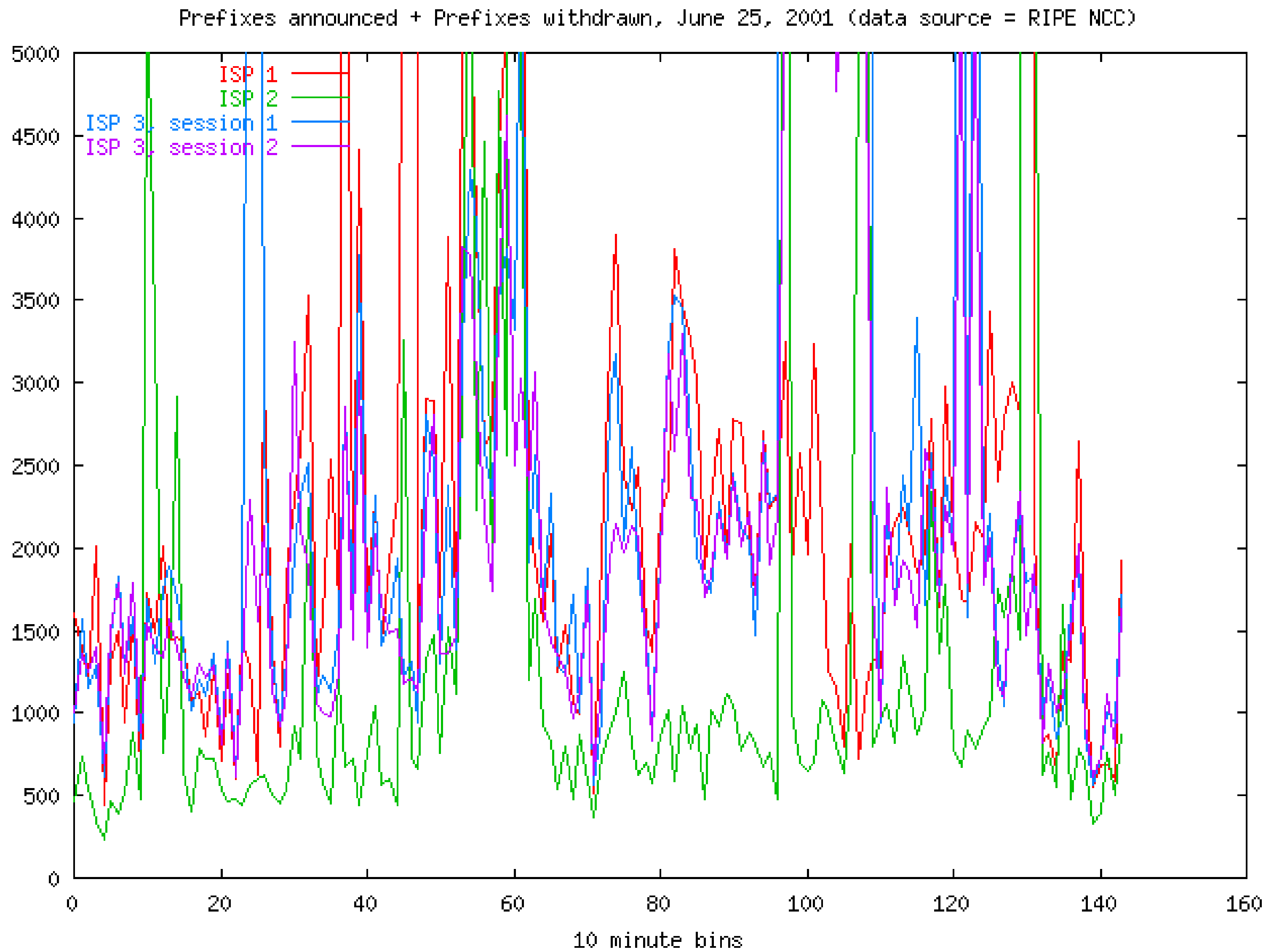
- How many updates are flying around the Internet?
- How long Does it take Routes to Change?

The goals of
(1) fast convergence
(2) minimal updates
(3) path redundancy
are at odds

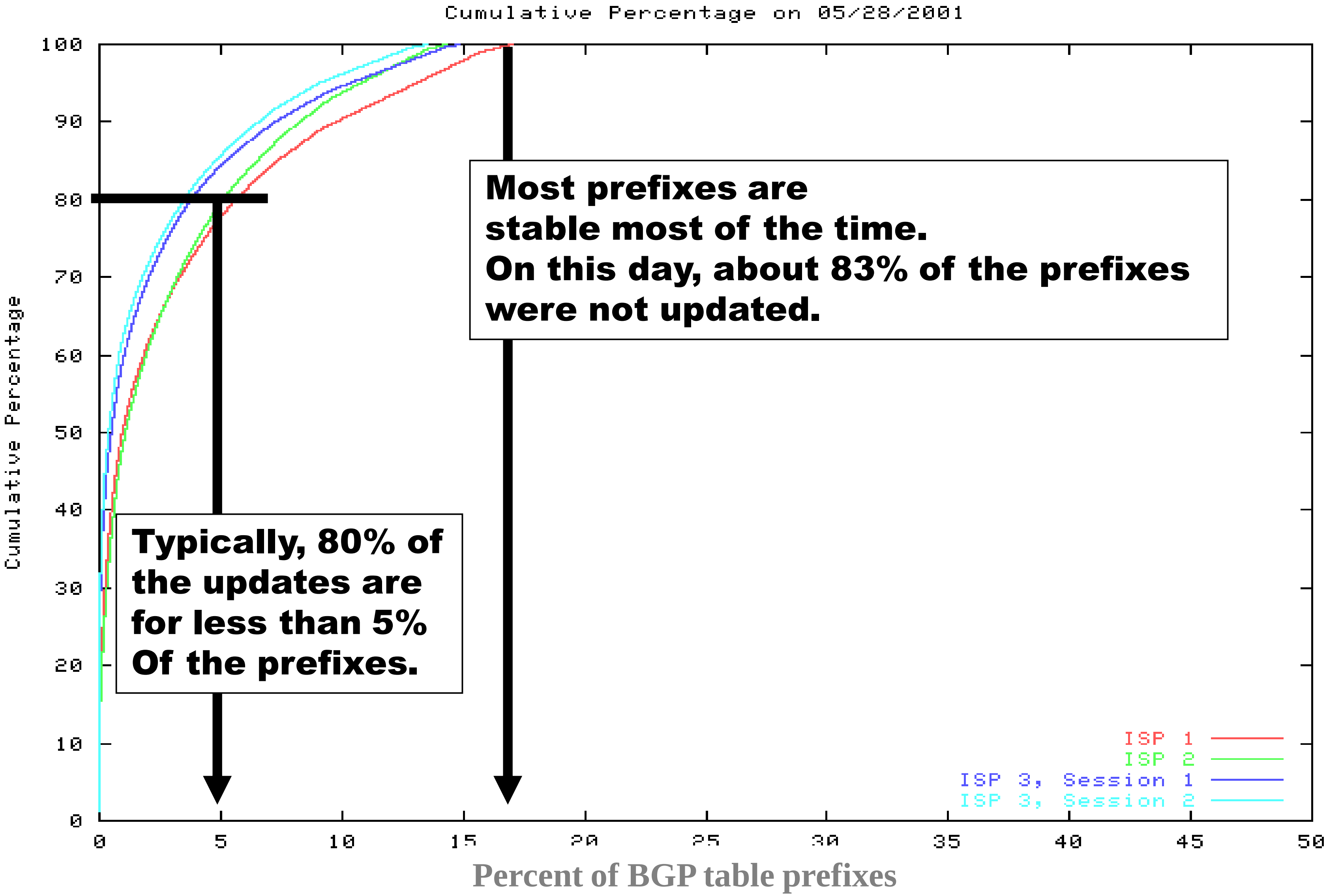
Daily Update Count



What is the Sound of One Route Flapping?



A Few Bad Apples ...



Thanks to Madanlal Musuvathi for this plot.

Data source: RIPE NCC

Two BGP Mechanisms for Squashing Updates

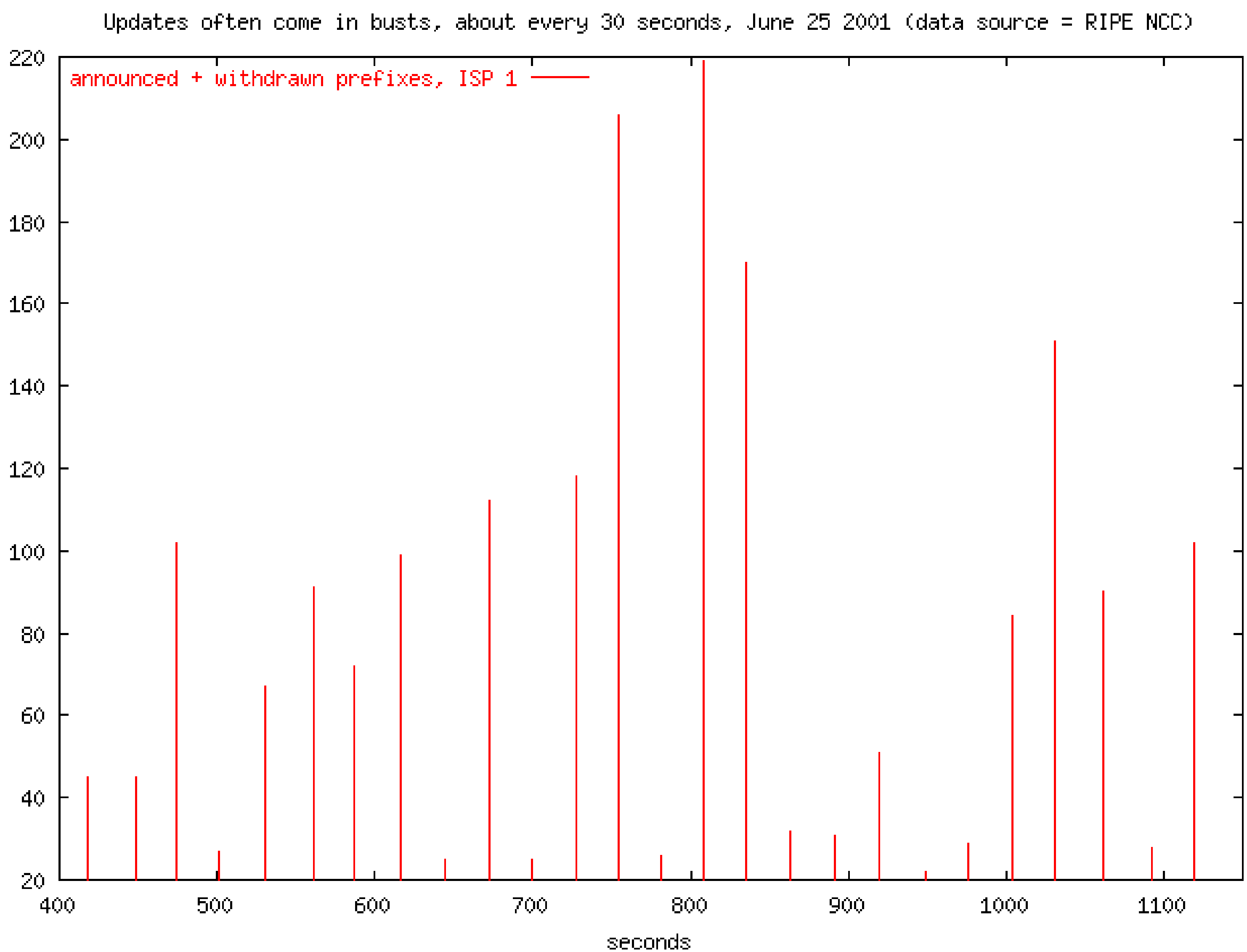
- **Rate limiting on sending updates**
  - Send batch of updates every MinRouteAdvertisementInterval seconds (+/- random fuzz)
  - Default value is 30 seconds
  - A router can change its mind about best routes many times within this interval without telling neighbors

Effective in dampening oscillations inherent in the vectoring approach

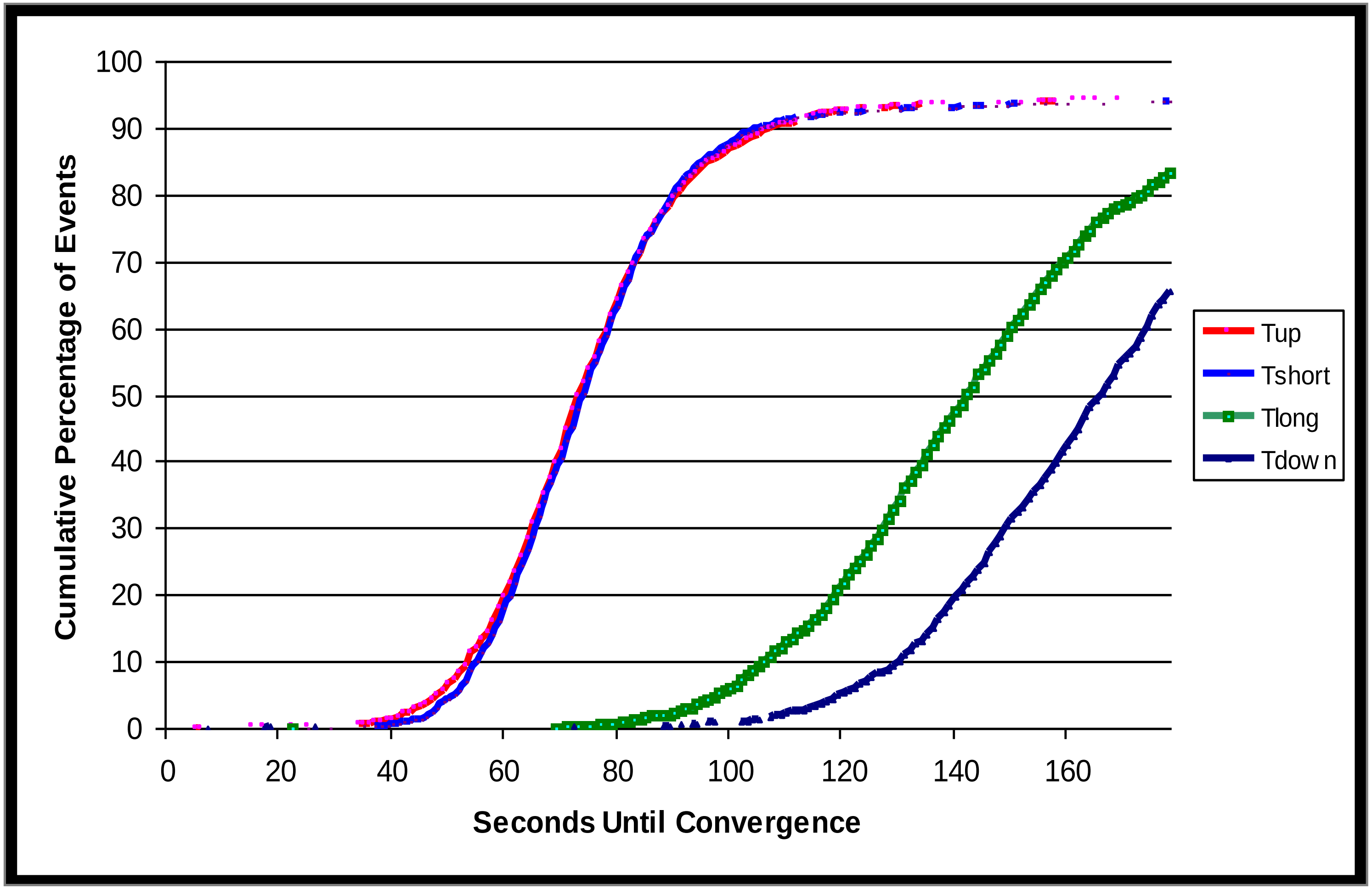
- **Route Flap Dampening**
  - Punish routes for “misbehaving”

Must be turned on with configuration

30 Second Bursts



How Long Does BGP Take to Adapt to Changes?



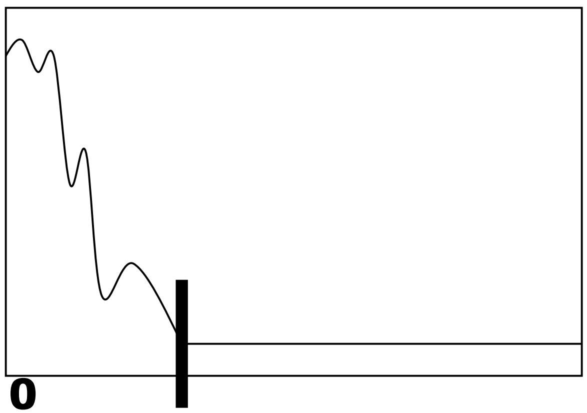
Thanks to Abha Ahuja and Craig Labovitz for this plot.

Two Main Factors in Delayed Convergence

- Rate limiting timer slows everything down
- BGP can explore many alternate paths before giving up or arriving at a new path
  - No global knowledge in vectoring protocols

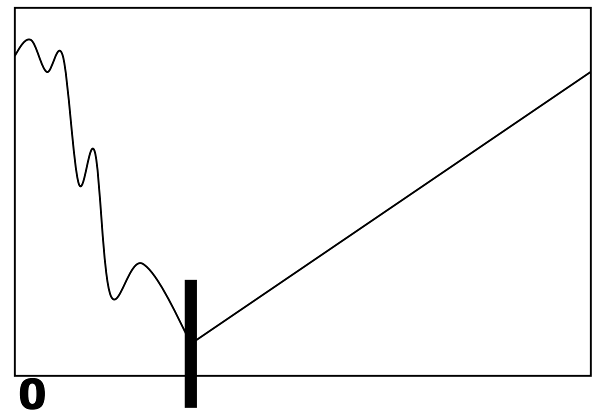
Why is Rate Limiting Needed?

Updates
to convergence



MinRouteAdvertisementInterval

Time
to convergence



MinRouteAdvertisementInterval

Rate limiting dampens some of the oscillation inherent in a vectoring protocol.

Current interval (30 seconds) was picked "out of the blue sky"

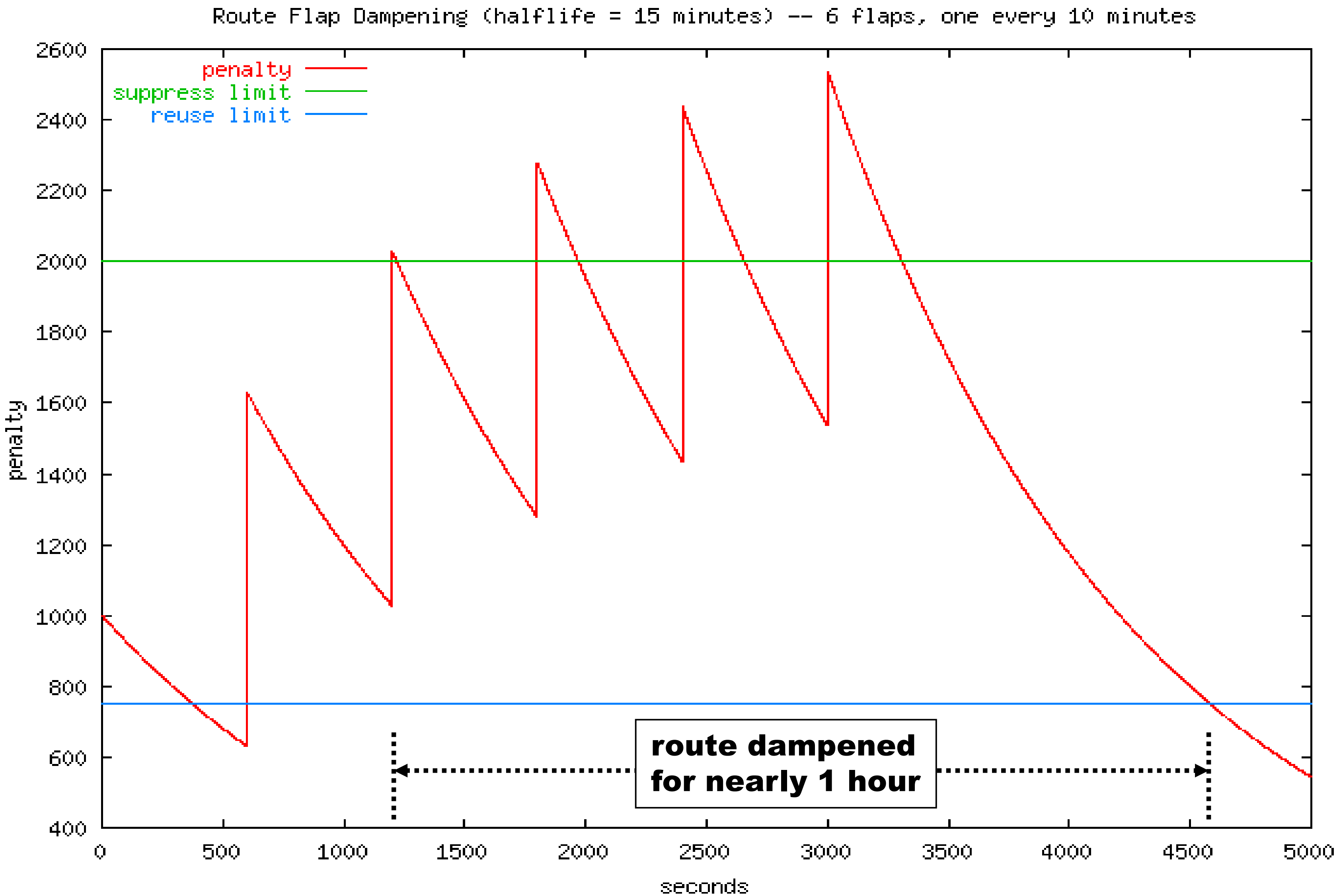
SSFNet (www.ssfnet.org) simulations, T. Griffin and B.J. Premore. To appear in ICNP 2001.

Route Flap Dampening (RFC 2439)

Routes are given a <u>penalty</u> for changing.
If penalty exceeds <u>suppress limit</u>, the
route is dampened. When the route is not changing,
its penalty decays exponentially. If the penalty goes
below <u>reuse limit</u>, then it is announced again.

- **Can dramatically reduce the number of BGP updates**
- **Requires additional router resources**
- **Applied on eBGP inbound only**

Route Flap Dampening Example



penalty for each flap = 1000

ICNP 2002

Q: Why All the Updates?

- Networks come, networks go
- There's always a router rebooting somewhere
- Hardware failure, flaky interface cards, backhoes digging, floods in Houston, …

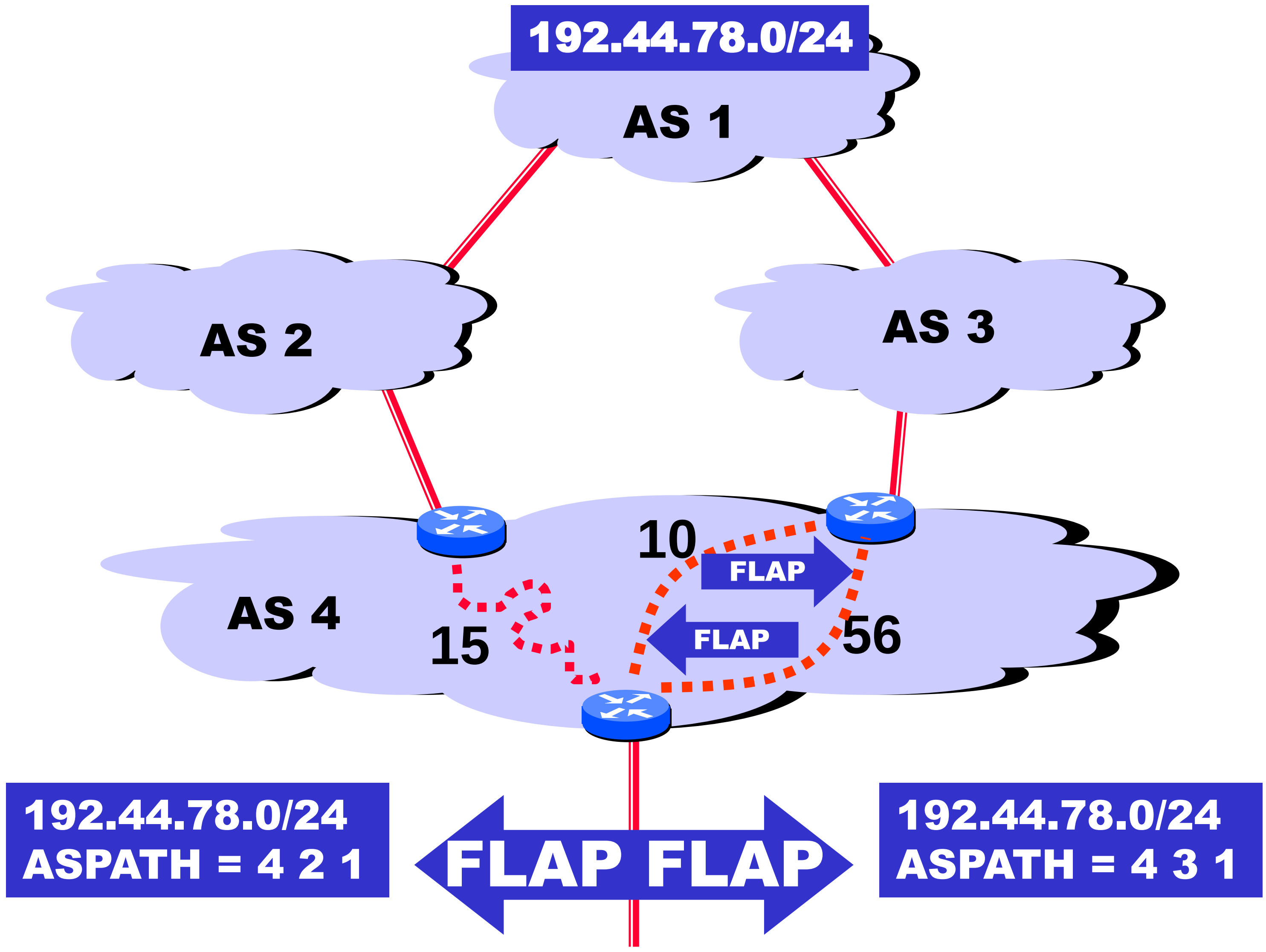
This is "normal" --- exactly what dynamic routing is designed for…

Q: Why All the Updates?

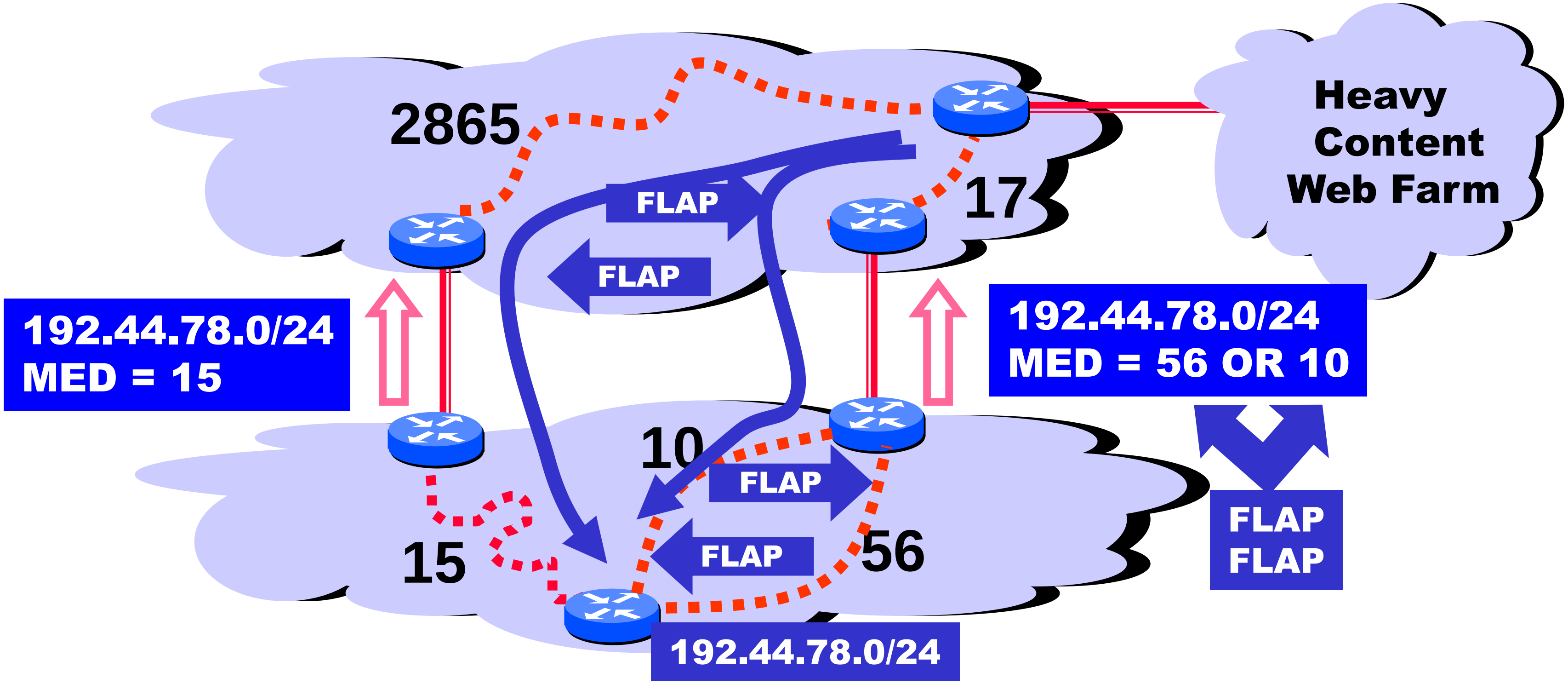
- Misconfiguration
- Route flap dampening not widely used
- BGP exploring many alternate paths
- Software bugs in implementation of routing protocols
- BGP session resets due to congestion or lack of interoperability: BGP sessions are brittle. One malformed update is enough to reset session and flap 100K routes. (Consequence of incremental approach)
- IGP instability exported by use of MEDs or IGP tie breaker
- Sub-optimal vendor implementation choices
- Secret sauce routing algorithms attempting fancy-dancy tricks
- Weird policy interactions (MED oscillation, BAD GADGETS??)
- Gnomes, sprites, and fairies
- ....

A: NO ONE REALLY KNOWS ...

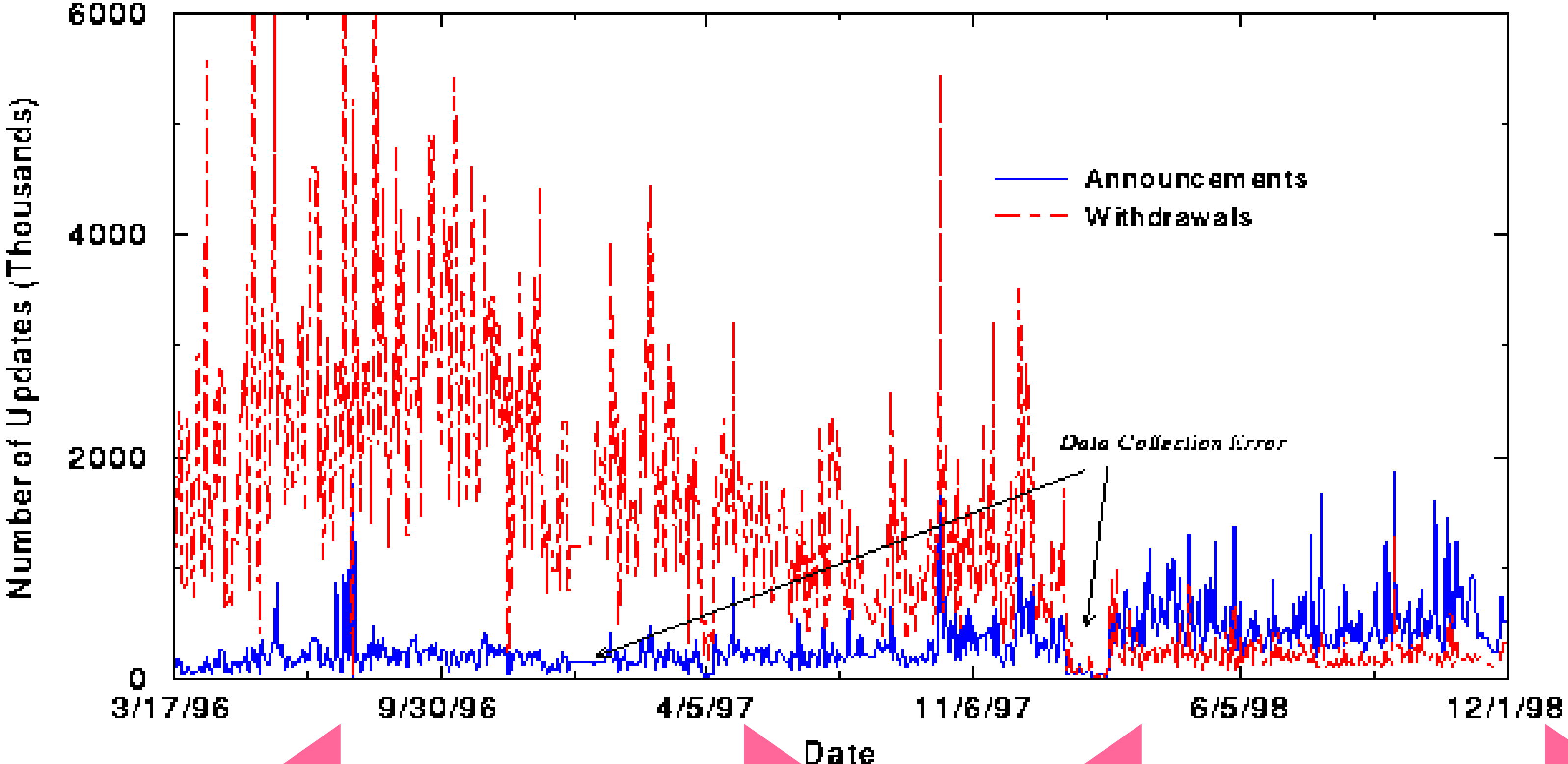
IGP Tie Breaking Can Export Internal Instability to the Whole Wide World



MEDs Can Export Internal Instability



Implementation Does Matter!



Thanks to Abha Ahuja and Craig Labovitz for this plot.

How Long Will Interdomain Routing Continue to Scale?

A quote from some recent email:

... the existing interdomain routing
infrastructure is rapidly nearing the
end of its useful lifetime. It
appears unlikely that mere tweaks
of BGP will stave off fundamental
scaling issues, brought on by growth,
multihoming and other causes.

Is this true or false? How can we tell?

Research required...

Summary

- BGP is a fairly simple protocol …
- … but it is not easy to configure
- BGP is running on more than 100K routers (my estimate), making it one of world's largest and most visible distributed systems
- Global dynamics and scaling principles are still not well understood

PART VII

Selected Bibliography

Addressing and ASN RFCs

- RFC 1380 IESG Deliberations on Routing and Addressing (1992)
- RFC 1517Applicability Statement for the Implementation of Classless Inter-Domain Routing (CIDR) (1993)
- RFC 1518 An Architecture for IP Address Allocation with CIDR (1993)
- RFC 1519 Classless Inter-Domain Routing (CIDR) (1993)
- RFC 1467 Status of CIDR Deployment in the Intrenet (1983)
- RFC 1520 Exchanging Routing Information Across Provider Boundaries in the CIDR Environment (1993)
- RFC 1817 CIDR and Classful routing (1995)
- RFC 1918 Address Allocation for Private Internets (1996)
- RFC 2008 Implications of Various Address Allocation Policies for Internet Routing (1996)
- RFC 2050 Internet Registry IP Allocation Guidelines (1996)
- RFC 2260 Scalable Support for Multi-homed Multi-provider Connectivity (1998)
- RFC 2519 A Framework for Inter-Domain Route Aggregation (1999)
- RFC 1930 Guidelines for creation, selection, and registration of an Autonomous System (AS)
- RFC 2270 Using a Dedicated AS for Sites Homed to a Single Provider

Selected BGP RFCs

Internet Engineering Task Force (IETF) http://www.ietf.org

- **IDR : http://www.ietf.org/html.charters/idr-charter.html**
- **RFC 1771 A Border Gateway Protocol 4 (BGP-4)**
  - **Latest draft rewrite: draft-ietf-idr-bgp4-18.txt**
- **RFC 1772 Application of the Border Gateway Protocol in the Internet**
- **RFC 1773 Experience with the BGP-4 protocol**
- **RFC 1774 BGP-4 Protocol Analysis**
- **RFC 2796 BGP Route Reflection An alternative to full mesh IBGP**
- **RFC 3065 Autonomous System Confederations for BGP**
- **RFC 1997 BGP Communities Attribute**
- **RFC 1998 An Application of the BGP Community Attribute in Multi-home Routing**
- **RFC 2439 Route Flap Dampening**

Titles of Some Recent Internet Drafts

- Dynamic Capability for BGP-4
- Application of Multiprotocol BGP-4 to IPv4 Multicast Routing
- Graceful Restart mechanism for BGP
- Cooperative Route Filtering Capability for BGP-4
- Address Prefix Based Outbound Route Filter for BGP-4
- Aspath Based Outbound Route Filter for BGP-4
- Architectural Requirements for Inter-Domain Routing in the Internet
- BGP support for four-octet AS number space
- Autonomous System Number Substitution on Egress
- BGP Extended Communities Attribute
- Controlling the redistribution of BGP routes
- BGP Persistent Route Oscillation Condition
- Benchmarking Methodology for Basic BGP Convergence
- Terminology for Benchmarking External Routing Convergence Measurements

BGP is a moving target ...

Selected Bibliography on Routing

- Internet Routing Architectures. Bassam Halabi. Second edition Cisco Press, 2000
- BGP4: Inter-domain Routing in the Internet. John W. Stewart, III. Addison-Wesley, 1999
- Routing in the Internet. Christian Huitema. 2000
- ISP Survival Guide: Strategies for Running a Competitive ISP. Geoff Huston. Wiley, 1999.
- Interconnection, Peering and Settlements. Geoff Huston. The Internet Protocol Journal. March and June 1999.

BGP Stability and Convergence

- Route Flap Damping Exacerbates Internet Routing Convergence. Z.M.Mao, R.Govindan, G.Varghese,R.H.Kranz. SIGCOMM 2002.
- The Impact of Internet Policy and Topology on Delayed Routing Convergence. Craig Labovitz, Abha Ahuja, Roger Wattenhofer, Srinivasan Venkatachary. INFOCOM 2001
- An Experimental Study of BGP Convergence. Craig Labovitz, Abha Ahuja, Abhijit Abose, Farnam Jahanian. SIGCOMM 2000
- Origins of Internet Routing Instability. C. Labovitz, R. Malan, F. Jahanian. INFOCOM 1999
- Internet Routing Instability. Craig Labovitz, G. Robert Malan and Farnam Jahanian. SIGCOMM 1997

Analysis of Interdomain Routing

- Cooperative Association for Internet Data Analysis (CAIDA)
  - http://www.caida.org/
  - Tools and analyses promoting the engineering and maintenance of a robust, scalable global Internet infrastructure
- Internet Performance Measurement and Analysis (IPMA)
  - http://www.merit.edu/ipma/
  - Studies the performance of networks and networking protocols in local and wide-area networks
- National Laboratory for Applied Network Research (NLANR)
  - http://www.nlanr.net/
  - Analysis, tools, visualization.
- IRTF Routing Research Group (IRTF-RR)
  - http://puck.nether.net/irtf-rr/
- Geoff Huston: http://bgp.potaroo.net

Internet Route Registries

- Internet Route Registry
  - http://www.irr.net/
- Routing Policy Specification Language (RPSL)
  - RFC 2622 Routing Policy Specification Language (RPSL)
  - RFC 2650 Using RPSL in Practice
- Internet Route Registry Daemon (IRRd)
  - http://www.irrd.net/
- RAToolSet
  - http://www.isi.edu/ra/RAToolSet/

Some BGP Theory

- Persistent Route Oscillations in Inter-Domain Routing. Kannan Varadhan, Ramesh Govindan, and Deborah Estrin. Computer Networks, Jan. 2000. (Also USC Tech Report, Feb. 1996)
  – Shows that BGP is not guaranteed to converge
- An Architecture for Stable, Analyzable Internet Routing. Ramesh Govindan, Cengiz Alaettinoglu, George Eddy, David Kessens, Satish Kumar, and WeeSan Lee. IEEE Network Magazine, Jan-Feb 1999.
  – Use RPSL to specify policies. Store them in registries. Use registry for conguration generation and analysis.
- An Analysis of BGP Convergence Properties. Timothy G. Griffin, Gordon Wilfong. SIGCOMM 1999
  – Model BGP, shows static analysis of divergence in policies is NP complete
- Policy Disputes in Path Vector Protocols. Timothy G. Griffin, F. Bruce Shepherd, Gordon Wilfong. ICNP 1999
  – Define Stable Paths Problem and develop sufficient condition for “sanity”
- A Safe Path Vector Protocol. Timothy G. Griffin, Gordon Wilfong. INFOCOM 2001
  – Dynamic solution for SPVP based on histories
- Stable Internet Routing without Global Coordination. Lixin Gao, Jennifer Rexford. SIGMETRICS 2000
  – Show that if certain guidelines are followed, then all is well.
- Inherently safe backup routing with BGP. Lixin Gao, Timothy G. Griffin, Jennifer Rexford. INFOCOM 2001
  – Use SPP to study complex backup policies

– On the Correctness of IBGP Configurations. Griffin and Wilfong.SIGCOMM 2002.
– An Analysis of the MED oscillation Problem. Griffin and Wilfong. ICNP 2002.